Exploring Grainyhead-like 2 target genes in breast cancer

Wang, Z.

Chapter 3

Genome-wide identification of binding sites of GRHL2 in luminal-like and basal A subtypes of breast cancer

Zi Wang[1], Haoyu Wu[2], Lucia Daxinger[2], Erik HJ Danen[1,3]

[1]Leiden Academic Center for Drug Research, Leiden University, Leiden, The Netherlands; [2]Department of Human Genetics, Leiden University Medical Centre, Leiden, The Netherlands; [3]correspondence to Erik HJ Danen, e.danen@lacdr.leidenuniv.nl

Abstract

Grainyhead like 2 (*GRHL2*) is one of three mammalian homologues of the Grainyhead (*GRH*) gene. It suppresses the oncogenic epithelial-mesenchymal transition (EMT), acting as a tumor suppressor. On the other hand, GHRL2 promotes cell proliferation by increasing human telomerase reverse transcriptase (hTERT) activity, serving as a tumor promoter. According to gene expression profiling, breast cancer can be divided into basal-like (basal A and basal B), luminal-like, HER2 enriched, claudin-low and normal-like subtypes. To identify common and subtype-specific genomic binding sites of GRHL2 in breast cancer, GRHL2 ChIP-seq was performed in three luminal-like and three basal A human breast cancer cell lines. Most binding sites of GRHL2 were found in intergenic and intron regions. 13,351 common binding sites were identified in basal A cells, which included 551 binding sites in gene promoter regions. For luminal-like cells, 6,527 common binding sites were identified, of which 208 binding sites were found in gene promoter regions. Basal A and luminal-like breast cancer cells shared 4711 GRHL2 binding sites, of which 171 binding sites were found in gene promoter regions. The identified GRHL2-binding motifs are all identical to a motif reported for human ovarian cancer, indicating conserved GRHL2 DNA-binding among human cancer cells. Notably, no binding sites of GRHL2 were detected in the promoter regions of several established EMT-related genes, including *CDH1*, *ZEB1*, *ZEB2* and *CDH2* genes. Collectively, this study provides a comprehensive overview of interactions of GRHL2 with DNA and lays the foundation for further understanding of common and subtype-specific signaling pathways regulated by GRHL2 in breast cancer.

Introduction

Breast cancer is the predominant cause of cancer-related death in women aged 20 to 59 years globally [1]. Based on gene expression profiling, breast cancer can be divided into several subtypes with distinct molecular features, which includes luminal-like (luminal A and luminal B), basal-like (basal A and basal B), human epidermal growth factor receptor 2 (HER2)-enriched, claudin-low and normal-like subtypes[2]. Both luminal-like and basal-like subtypes comprise at least 73% of all breast cancers[2]. Conversions of luminal to basal lineage have been observed in mouse breast cancer models [3,4] but luminal-like and basal-like subtypes differ in prognosis and response to therapy. Therefore, it is important to characterize common features and discordances between them.

The *GRH* gene was discovered in *Drosophila* and its mammalian homologs have three members (*GRHL1*, *GRHL2* and *GRHL3*)[5]. *GRH* deficiency leads to failure of complete neural tube closure, epidermal barrier formation, trachea elongation and epidermal wound response [5-7]. *GRHL2* is one of three mammalian homologues of the *GRH* gene, which has been investigated in cancer development. *GRHL2* is located on chromosome 8q22 that is frequently amplified in many cancers, including breast cancer, colorectal cancer and oral squamous cell carcinoma[8-10]. GRHL2, as an oncogene, positively regulates cell proliferation by enhancing hTERT activity through inhibition of DNA methylation at 5'-CpG island around gene promoter[9]. GRHL2 inhibits cell apoptosis by suppressing death receptor (FAS and DR5) expression in breast cancer cells[8,11]. Knockdown of GRHL2 downregulated HER3 expression, resulting in inhibition of cell proliferation[11]. On the other hand, GRHL2 was previously reported as a suppressor of oncogenic EMT by the loop of GRHL2-miR200-ZEB1 and regulation of the TGF-β pathway[12-14]. These controversial results suggest that the roles of GRHL2 may be tumor-specific through regulating different target genes in different cancers.

Chromatin immunoprecipitation followed by deep sequencing (ChIP-seq) is a widely used method to analyze protein-DNA interactions, histone modifications, and nucleosomes on genome-wide scale in living cells by capturing proteins at sites of their binding to DNA[15,16]. Previous findings showed that GRHL2 shares a similar DNA-binding motif with other GRHL family members[13,17,18]. To date, no studies have

investigated the genomic landscape of GRHL2 binding sites across breast cancer subtypes. In this study, we provide a comprehensive overview of binding sites of GRHL2 in the genome of basal A and luminal-like subtypes of breast cancer.

Methods and materials

Cell lines

Human breast cancer cell lines representing luminal-like (MCF7, T47D, BT474), basal A (HCC1806, BT20 and MDA-MB-468), and basal B subtypes (Hs578T) were obtained from the American Type Culture Collection. Cells were cultured in RPMI1640 medium with 10% fetal bovine serum, 25 U/mL penicillin and 25 µg/mL streptomycin in the incubator (37°C, 5% CO2).

Chromatin immunoprecipitation-sequencing (ChIP-seq)

Cells were grown in RPMI-1640 complete medium. Cross-linking was performed by 1% formaldehyde for 10 minutes at room temperature (RT). Then 1M glycine (141 µl of 1M glycine for 1 ml of medium) was used to quench for 5 minutes at RT. Cells were washed twice with ice-cold PBS containing 5 µl/ml phenylmethylsulfonyl fluoride (PMSF). Cells were harvested by centrifugation (2095 g for 5 minutes at 4°C) and lysed with NP40 buffer (150 mM NaCl, 50mM Tris-HCl, 5mM EDTA, 0.5% NP40, 1% Triton X-100) containing 0.1% SDS, 0.5% sodium deoxycholate and protease inhibitor cocktail (EDTA-free Protease Inhibitor Cocktail, Sigma). Chromatin was sonicated to an average size of 300 bp (Fig. S1). GRHL2-bound chromatin fragments were immunoprecipitated with anti-GRHL2 antibody (Sigma; HPA004820). Precipitates were eluted by NP buffer, low salt (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl (pH 8.1), 150mM NaCl), high salt (0.1% SDS, 1% Triton X-100, 2mM EDTA, 20mM Tris-HCl (pH 8.1), 500mM NaCl) and LiCl buffer (0.25M LiCl, 1%NP40, 1% deoxycholate, 1mM EDTA, 10mM Tris-HCl (pH 8.1)). Chromatin was de-crosslinked by 1% SDS at 65°C. DNA was purified by Phenol:Chloroform:Isoamyl Alcohol (PCI) and then diluted in TE buffer.

In order to examine the quality of our samples before sequencing, ChIP-PCR was performed to validate interaction of GRHL2 with the promoter region of *CLDN4*, a direct

target gene of GRHL2 [19]. The results confirmed the GRHL2 binding site around the *CLDN4* promoter (Fig. S2). The following primers were used for ChIP-PCR: *CLDN4* forward: gtgacctcagcatgggctttga, *CLDN4* reverse: ctcctcctgaccagtttctctg, Control (an intergenic region upstream of the *GAPDH* locus) forward: atgggtgccactggggatct, Control reverse: tgccaaagcctaggggaaga, *ZEB1* promoter# forward: cggtccctagcaacaaggtt, *ZEB1* promoter# reverse: tcgcttgtgtctaaatgctcg. *ZEB1*## forward: gccgccgagcctccaacttt, *ZEB1*## reverse: tgctagggaccgggcggttt, *OVOL2* exon forward: ccttaaatcgcgagtgagacc, *OVOL2* exon reverse: gtagcgagcttgttgacacc, *CDH1* intron forward: gtatgaacggcaagcctctg, *CDH1* intron reverse: caagggagccaggaagagaa. ChIP-PCR data were collected and analyzed using the $2^{-\Delta\Delta Ct}$ method [20].

For ChIP-Seq, library preparation and paired-end sequencing were performed by GenomeScan (Leiden, The Netherlands)

ChIP-seq data analysis

Paired-end reads were mapped to the human reference genome (hg38) using BWA-MEM[21] with default parameters. Over 93% of total reads were mapped to the human genome in BT20, HCC1806, MDA-MB-468, T47D and MCF7 cell line. For BT474, ~57.3% reads were mapped. Phred quality score (Q score) was used to measure base calling accuracy, which indicates the probability that a given base is called incorrectly[22]. Q score is logarithmically related to the base calling error probabilities *P* [22].

$$Q = -10 \log_{10} P$$

Q=30 nominally corresponds to a 0.1% error rate[23]. Reads with scores > Q30 were over 86% in BT20, HCC1806, MDA-MB-468, T47D and MCF7 cell lines. For BT474, reads with scores > Q30 accounted for 48.6%.

To examine whether the paired-end reads were appended with unwanted adapter sequences, an adapter content test was performed. The quality control report (Fig. S3) showed that cumulative presence of adapter sequences was <5% in all cell samples, indicating that all data sets could be further analyzed without adapter-trimming. Per base sequence quality of sequencing was examined, which indicated that all sequencing data were of high quality (Fig. S4) and could be further analyzed.

Reads with low mapping quality (≤ Q30) were filtered out. MACS version 2.1.0[24] was used for peak calling by default settings. q value was adjusted to 0.1 for BT474 cell line to avoid loss of peaks. The annotatePeaks and MergePeaks function from HOMER [25] were used to annotate and overlap peaks, respectively. ChIPseeker was used for analysis of ChIP-seq peaks coverage plot and density profile of GRHL2 binding sites [26]. Motif analysis was performed by ChIP-seq peaks with high scores using the MEME-ChIP program with default settings. ChIP-seq data was visualized by UCSC genome browser.

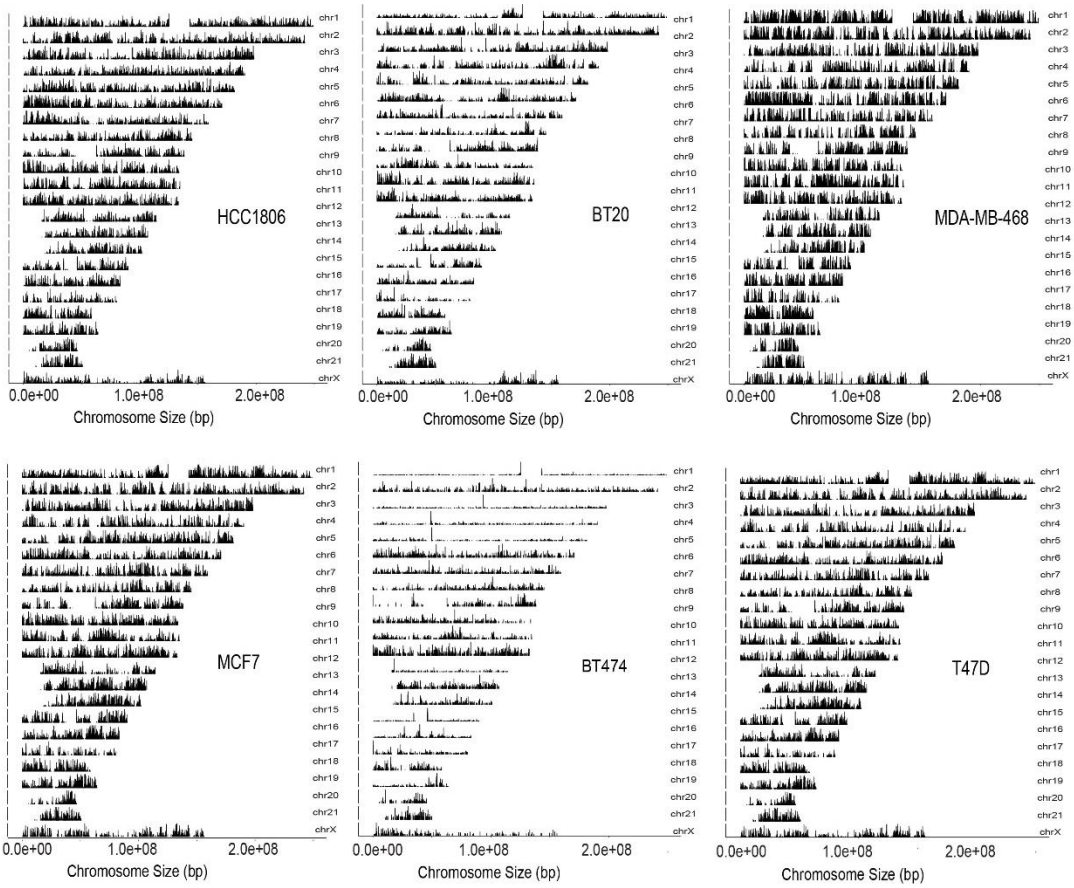


Fig. 1 Coverage of peak regions across chromosomes. The graph represents the coverage of GRHL2 binding sites across the chromosomes**.**

Results

Genome-wide identification of binding sites of GRHL2 in luminal-like and basal A subtypes of breast cancer

To identify GRHL2 binding sites, ChIP-seq was performed in luminal-like (MCF7, T47D and BT474) and basal A (HCC1806, BT20 and MDA-MB-468) breast cancer cells.

Firstly, the coverage of peak regions across chromosomes was analyzed[26]. In each cell sample, GRHL2 was strongly associated with all chromosomes (Fig. 1).

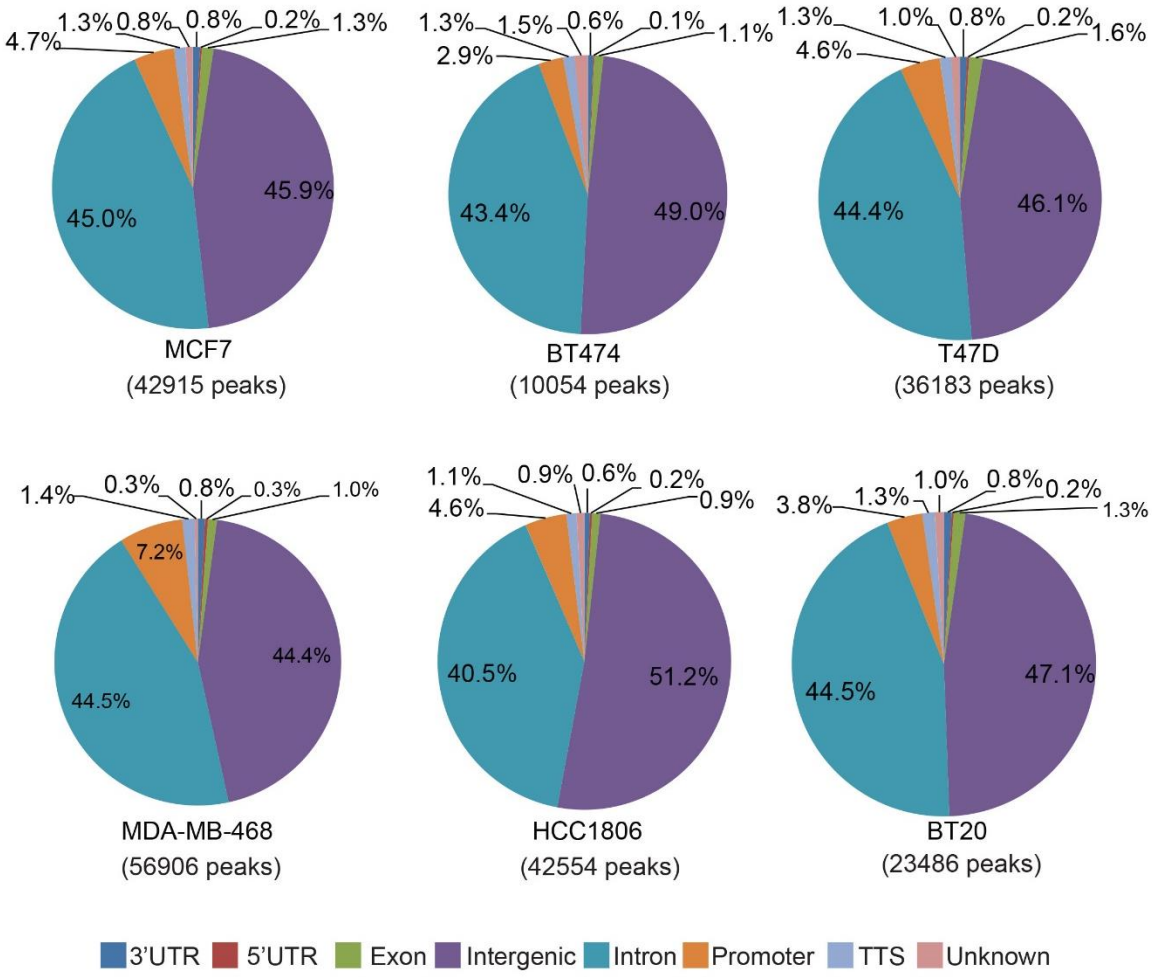


Fig. 2 Percentage of GRHL2 binding sites found at promoter regions, 5' untranslated regions (UTRs), 3' UTRs, exons, introns, intergenic regions, transcription termination sites (TTSs) and unknown regions. Promoter regions are defined as -1000 bp to +100 bp from the transcription start sites (TSS).

GRHL2 binding sites were found in intergenic regions, transcription start sites (TSS) promoter regions, introns, exons, transcription termination sites (TTS) and unknown regions (Fig. 2). The majority of peaks was located in intergenic and intron regions in basal A and luminal-like breast cancer cells. Genes where GRHL2 was found to interact with the -1000 bp to +100 bp promoter region in all three luminal (left column), all three basal A (middle column), or all luminal and basal A cell lines tested (right column) were identified and represent likely candidate general and subtype-specific GRHL2 target genes (Table S1).

To further investigate if peaks were enriched in promoter regions, read count frequency and density profiling of GRHL2 binding sites within -6000 bp ~ +6000 bp of the transcription start site (TSS) were analyzed (Fig. 3). Consistent with the annotation of binding sites, which showed most GRHL2 binding sites existed in the intergenic regions, the density of GRHL2 binding sites was not increased in the -1000 bp to +100 bp promoter region of basal A and luminal-like breast cancer cells.

To detect similarities of GRHL2 binding sites between luminal-like and basal A subtype, three luminal-like/basal A data sets were overlapped to identify shared binding sites. 13,351 common binding sites were identified in basal A subtype of breast cancer cells, which included 551 binding sites in gene promoter regions (-1000 bp~ +100 bp from TSS) (Fig. 4a and b). For luminal-like breast cancer cells, 6,527 common binding sites were identified, of which 208 binding sites were found in gene promoter regions (Fig. 4c and d). Basal A and luminal-like subtypes of breast cancer cells shared 4,711 binding sites of GRHL2, of which 171 binding sites were found in gene promoter regions (Fig. 4e and f).

Identification of a common GRHL2-interaction motif

The MEME-ChIP program was used to identify motifs, all of which were with statistical significance. In each sample, 3 motifs with high E value were shown (Fig. 5), whose core binding was similar to previously published ones [13,27-29]. Thus, our ChIP-seq data indicated that GRHL2 motif was highly conserved in human and mouse cells.

GRHL2-binding at EMT-related genes

GRHL2 and OVOL2 support an epithelial phenotype and counteract EMT transcription factor such as ZEB and SNAIL. Some studies have reported that GRHL2 binding sites are present in the intronic region of *CDH1* and in the promoter regions of *CLDN4* and *OVOL2* for activation of transcription and GRHL2 was reported to bind the *ZEB1* gene as a negative regulator [12,27,30,31]. In our ChIP-seq data, GRHL2 binding sites were observed at *CDH1* introns and at promoter regions of *CLDN4* and *OVOL2* (Fig. 6) ChIP-PCR was performed to further validate these interactions (Fig. 7). *CLDN4* showed multiple GRHL2 binding sites across the coding and non-coding regions,

suggesting the binding of GRHL2 to multiple regions may be involved in long-distance chromatin interactions as suggested previously [13]. Conversely, no GRHL2 binding was observed at the promoter of *ZEB1* or *ZEB2* (Fig. 6), arguing against mutual regulation through direct interaction as previously suggested [32,33]. To further evaluate this, ChIP-PCR was carried out using primers that have been previously reported to amplify *ZEB1* promoter DNA sequences bound by GRHL2 in human mammary epithelial cells and human ovarian cancer cells. This experiment further confirmed the absence of GRHL2 binding sites around the promoter of *ZEB1* in basal A (HCC1806, BT20, MDA-MB-468) and luminal-like (T47D, BT474) subtype breast cancer cells (Fig. 7). Moreover, interactions of GRHL2 with *CDH1* intron and *OVOL2* promoter regions were validated in these experiments (Fig. 7). Together, these findings suggest that GRHL2 binding sites in EMT-related genes may be cell context-dependent.

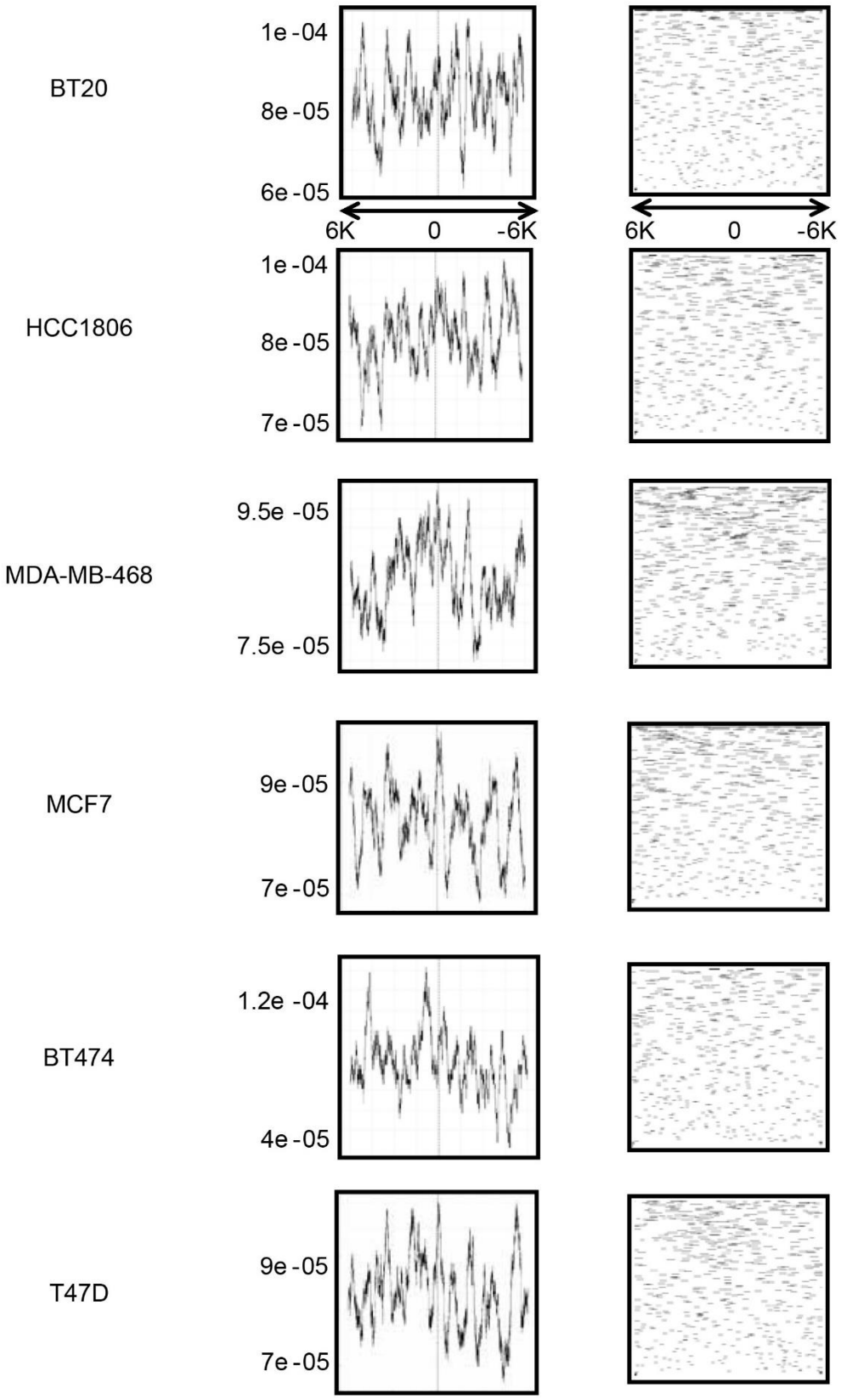


Fig. 3 The read count frequency and density profile of GRHL2 binding sites within -6000 bp~ +6000 bp of the promoter-TSSs. On the left side, graphs are for GRHL2 ChIP-seq read count frequency in indicated cell line. X axis represents read count frequency; Y axis is for genomic

region. On the right side, graphs show the density of ChIP-seq reads for GRHL2 binding sites in the indicated cell line.

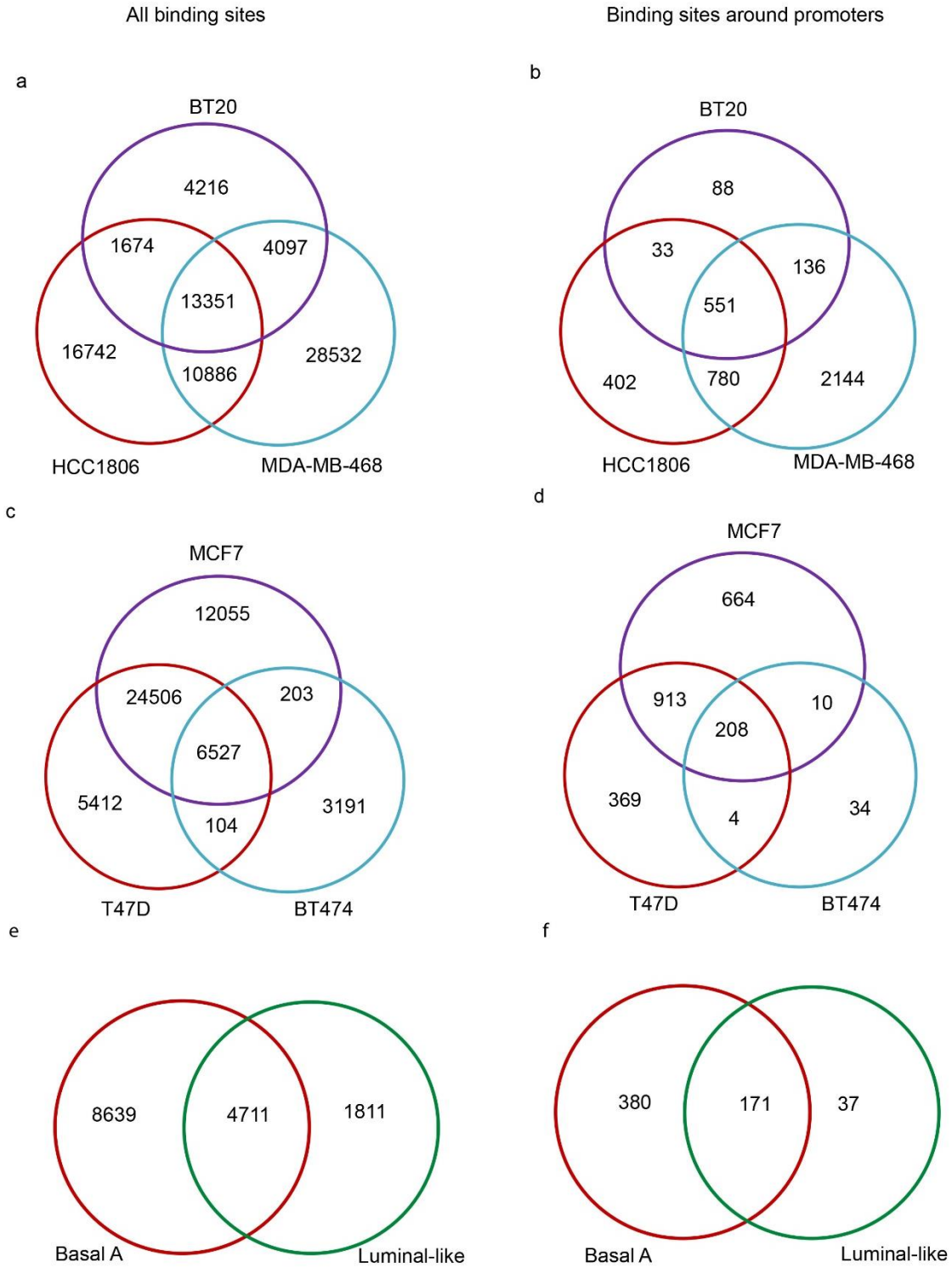


Fig. 4 Overlap of GRHL2 binding sites. Overlap of GRHL2 binding sites is identified in the indicated subtypes.

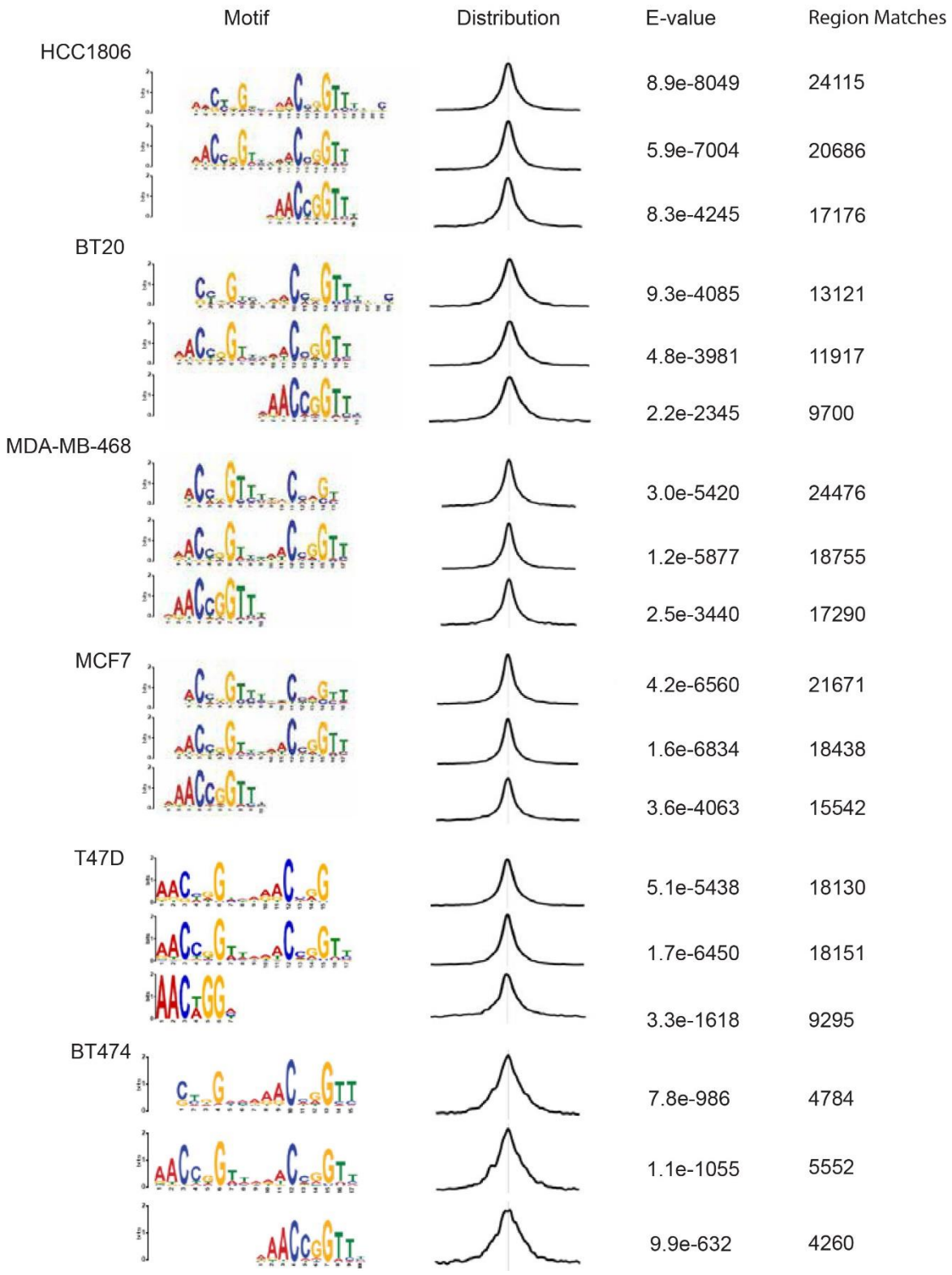


Fig. 5 DNA-binding motif of GRHL2 in luminal-like and basal A subtypes of breast cancer. From left to right, the first panel shows the identified motifs in the indicated cells. The second panel shows distribution of the best matches to the motif in the sequences. The third panel shows E-value, the significance of the motif according to the motif discovery. The last panel shows the number of regions that match the corresponding motif.

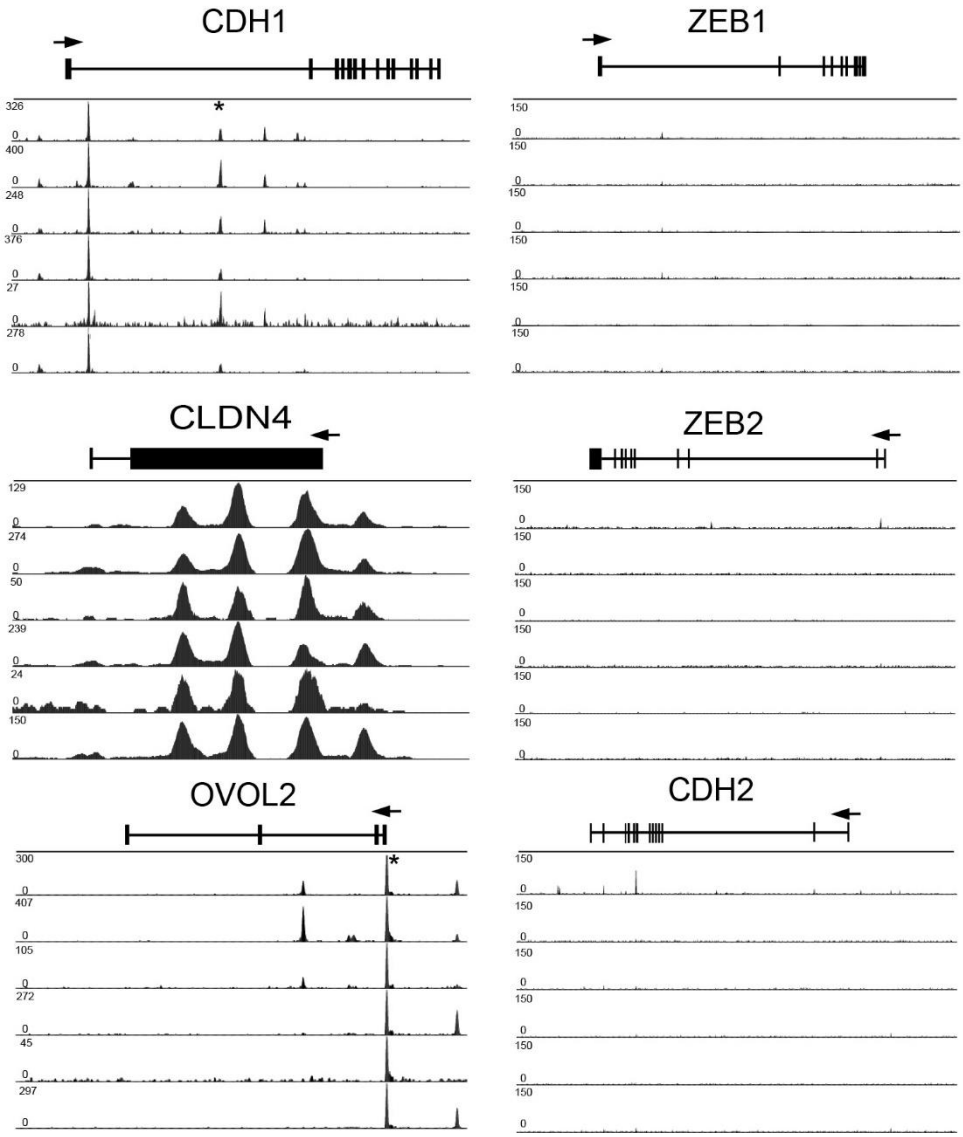
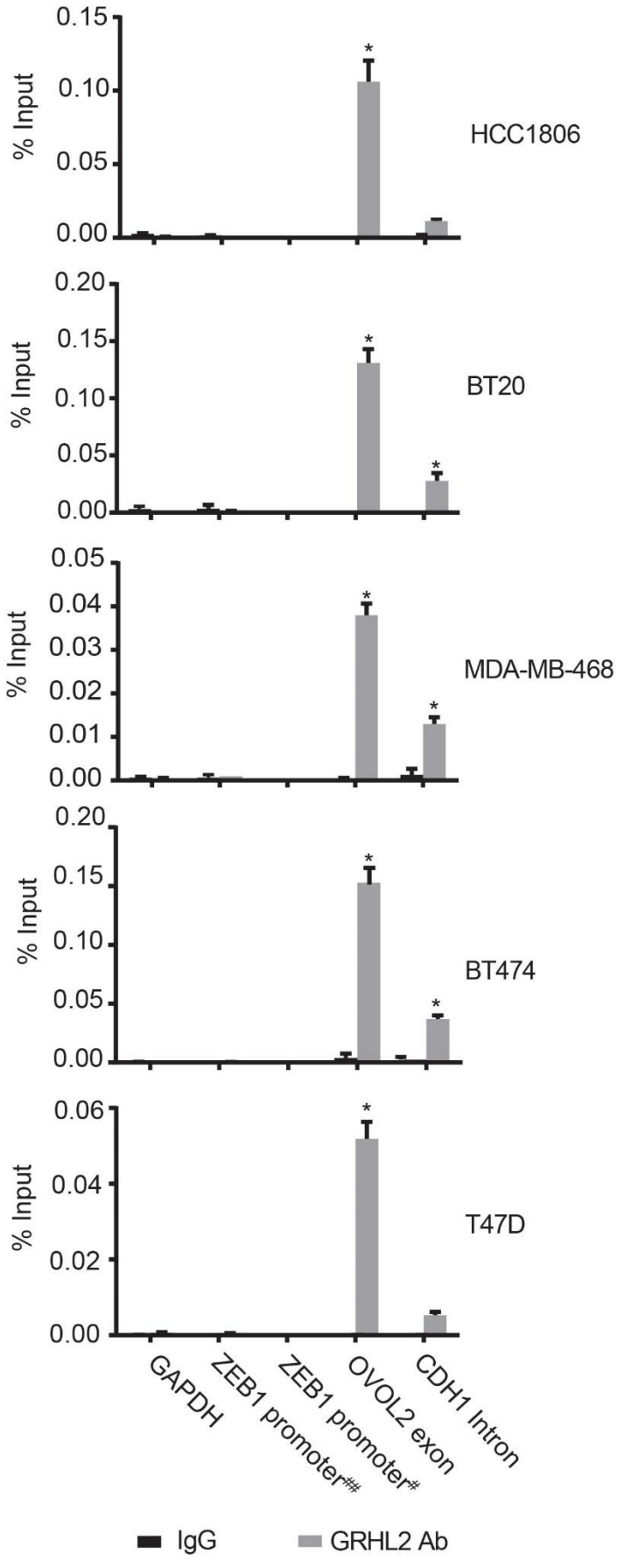


Fig. 6 GRHL2 ChIP tracks at selected core genes and EMT genes. ChIP tracks are shown from top to bottom for HCC1806, MDA-MB-468, BT20, MCF7, BT474 and T47D, respectively. The snapshot on the left shows results for 3 identified GRHL2 targets (*CDH1*, *CLDN4* and *OVOL2*) and the snapshot on the right shows results for three not-identified genes encoding proteins associated with EMT (*ZEB1*, *ZEB2* and *CDH2*). The track height is scaled from 0 to the indicated number. Above all tracks, the locus with its exon/intron structure is presented. Binding sites with * are validated by ChIP-PCR.


0.15
0.10
0.05
0.00
% Input
HCC1806
0.20
0.15
0.10
0.05
0.00
% Input
BT20
0.05
0.04
0.03
0.02
0.01
0.00
% Input
MDA-MB-468
0.20
0.15
0.10
0.05
0.00
% Input
BT474
0.06
0.04
0.02
0.00
% Input
T47D
GAPDH
ZEB1 promoter##
ZEB1 promoter#
OVOL2 exon
CDH1 Intron
IgG
GRHL2 Ab

(Last page) Fig. 7. ChIP-PCR validation of presence and absence of GRHL2 binding sites identified by ChIP-seq. Graphs represent the efficiency of indicated genomic DNA co-precipitation with anti-GRHL2 Ab (black bars) or IgG control Ab (grey bars). ChIP-PCR showing enrichment of GRHL2 binding sites at *OVOL2* exon and *CDH1* intron, but not *ZEB1* promoter regions. For *ZEB1* detection, primers that were previously reported to successfully amplify GRHL2 binding sites in human mammary epithelial cells (##) and human ovarian cancer cells (#) were used. Signals for IgG control and GRHL2 antibody pulldown samples are normalized to input DNA and are presented as % input with SEM from 3 technical replicates. Data are statistically analyzed by t-test and * indicates $p < 0.05$.

Discussion

Cell type origin is one of the most important factors that determine molecular features of tumors [34]. In general, luminal-like tumor cells are biologically similar to cells derived from inner (luminal) cells lining the mammary ducts, whereas cells of basal-like breast cancer are characterized by features similar to surrounding the mammary ducts[35]. Basal-like breast cancers are associated with a worse prognosis and an increased possibility of cancer metastasis compared with the luminal-like subtype[4,36,37]. Immunohistochemical staining is clinically used to categorize luminal-like breast cancer into luminal A (estrogen receptor (ER) and/or progesterone receptor (PR) positive, HER2 negative) and luminal B (ER and/or PR, and HER2 positive). However, most basal-like breast cancers are negative for ER, PR and HER2, therefore the majority of basal-like breast cancer is triple negative breast cancer (TNBC). Basal-like breast cancer can be further subdivided into basal A and basal B. As for basal A, it is associated with *BRCA1* signatures and resembles basal-like tumors, whereas basal B subtype displays mesenchymal properties and stem/progenitor characteristics[38,39].

In the present study, ChIP-seq was performed to characterize genome-wide binding sites of transcription factor GRHL2 in basal A and luminal-like subtypes of breast cancer. The match with previously a published binding motif shows that GRHL2-interaction with the DNA is highly conserved in human cancer cells. A limited number of binding sites were located in gene promoter regions. Similar to previous reports [13,28], most binding sites were located in introns and intergenic regions of target genes. Such regions may contain enhancers interacting with GRHL2 and GRHL2 has also been reported to regulate histone modifications such as H3K4me3 and H3K4me1[13,40]. Together, this suggests that GRHL2 may regulate gene expression through direct

transcriptional control at promoter regions or through alternative mechanisms including epigenetic mechanisms.

Close to 5000 identified GRHL2 genomic binding sites were shared between all tested basal A and luminal-like cell lines. A similar number of binding sites were found in all basal-like cell lines but were not detected in any of the tested luminal lines. These candidate subtype-specific GRHL2-target sites may serve as a starting point to the unraveling of distinct transcriptional networks in different breast cancer subtypes.

Our analysis of GRHL2 interaction with known EMT-related genes fits previously published findings except for *ZEB1*. It was reported that ZEB1 is regulated by GRHL2 directly and, vice versa, that ZEB1 regulates GRHL2 in a balance between EMT and MET [10-12,32]. However, we did not detect obvious GRHL2 binding sites in the promoter regions of the *ZEB1* or *ZEB2* genes. GRHL2 may regulate *ZEB1* and *ZEB2* indirectly in luminal-like and basal A breast cancers.

Taken together, this study provides a comprehensive genome-wide resource of GRHL2 binding sites and identifies specific and shared binding sites for GRHL2 in luminal-like and basal A subtype breast cancer. Overall, this study lays the foundation for unraveling signaling pathways regulated by GRHL2.

Acknowledgements

Zi Wang was supported by the China Scholarship Council. This work was supported by the Dutch Cancer Society (KWF Research Grant #10967).

References

1 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA: a cancer journal for clinicians* **66**, 7-30 (2016).

2 Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol* **5**, 412-424, doi:10.5306/wjco.v5.i3.412 (2014).

3 Cheung, K. J., Gabrielson, E., Werb, Z. & Ewald, A. J. Collective invasion in breast cancer requires a conserved basal epithelial program. *Cell* **155**, 1639-1651, doi:10.1016/j.cell.2013.11.029 (2013).

4 Sonzogni, O. *et al.* Reporters to mark and eliminate basal or luminal epithelial cells in culture and in vivo. *PLoS biology* **16**, e2004049, doi:10.1371/journal.pbio.2004049 (2018).

5 Frisch, S. M., Farris, J. C. & Pifer, P. M. Roles of Grainyhead-like transcription factors in cancer. *Oncogene*, doi:10.1038/onc.2017.178 (2017).

6 Bray, S. J. & Kafatos, F. C. Developmental function of Elf-1: an essential transcription factor during embryogenesis in Drosophila. *Genes & development* **5**, 1672-1683 (1991).

7 Mace, K. A., Pearson, J. C. & McGinnis, W. An epidermal barrier wound repair pathway in Drosophila is mediated by grainy head. *Science* **308**, 381-385, doi:10.1126/science.1107573 (2005).

8 Dompe, N. *et al.* A whole-genome RNAi screen identifies an 8q22 gene cluster that inhibits death receptor-mediated apoptosis. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E943-951, doi:10.1073/pnas.1100132108 (2011).

9 Chen, W. *et al.* Grainyhead-like 2 enhances the human telomerase reverse transcriptase gene expression by inhibiting DNA methylation at the 5'-CpG island in normal human keratinocytes. *The Journal of biological chemistry* **285**, 40852-40863, doi:10.1074/jbc.M110.103812 (2010).

10 Quan, Y. *et al.* Downregulation of GRHL2 inhibits the proliferation of colorectal cancer cells by targeting ZEB1. *Cancer Biol Ther* **15**, 878-887, doi:10.4161/cbt.28877 (2014).

11 Werner, S. *et al.* Dual roles of the transcription factor grainyhead-like 2 (GRHL2) in breast cancer. *The Journal of biological chemistry* **288**, 22993-23008, doi:10.1074/jbc.M113.456293 (2013).

12 Cieply, B., Farris, J., Denvir, J., Ford, H. L. & Frisch, S. M. Epithelial-mesenchymal transition and tumor suppression are controlled by a reciprocal feedback loop between ZEB1 and Grainyhead-like-2. *Cancer Res* **73**, 6299-6309, doi:10.1158/0008-5472.CAN-12-4082 (2013).

13 Chung, V. Y. *et al.* GRHL2-miR-200-ZEB1 maintains the epithelial status of ovarian cancer through transcriptional regulation and histone modification. *Sci Rep* **6**, 19943, doi:10.1038/srep19943 (2016).

14 Gregory, P. A. *et al.* An autocrine TGF-beta/ZEB/miR-200 signaling network regulates establishment and maintenance of epithelial-mesenchymal transition. *Mol Biol Cell* **22**, 1686-1698, doi:10.1091/mbc.E11-02-0103 (2011).

15 Satoh, J.-i., Kawana, N. & Yamamoto, Y. Pathway analysis of ChIP-Seq-based NRF1 target genes suggests a logical hypothesis of their involvement in the pathogenesis of neurodegenerative diseases. *Gene regulation and systems biology* **7**, 139 (2013).

16 Mundade, R., Ozer, H. G., Wei, H., Prabhu, L. & Lu, T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle* **13**, 2847-2852, doi:10.4161/15384101.2014.949201 (2014).

17 Wilanowski, T. *et al.* A highly conserved novel family of mammalian developmental transcription factors related to Drosophila grainyhead. *Mechanisms of development* **114**, 37-50 (2002).

18 Boglev, Y. *et al.* The unique and cooperative roles of the Grainy head-like transcription factors in epidermal development reflect unexpected target gene specificity. *Developmental biology* **349**, 512-522, doi:10.1016/j.ydbio.2010.11.011 (2011).

19 Werth, M. *et al.* The transcription factor grainyhead-like 2 regulates the molecular composition of the epithelial apical junctional complex. *Development* **137**, 3835-3845, doi:10.1242/dev.055483 (2010).

20 Lin, X., Tirichine, L. & Bowler, C. Protocol: Chromatin immunoprecipitation (ChIP) methodology to investigate histone modifications in two model diatom species. *Plant Methods* **8**, 48, doi:10.1186/1746-4811-8-48 (2012).

21 Liu, C.-M. *et al.* SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. **28**, 878-879 (2012).

22 Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175-185 (1998).

23 Liao, P., Satten, G. A. & Hu, Y. J. PhredEM: a phred-score-informed genotype-calling approach for next-generation sequencing studies. *Genet Epidemiol* **41**, 375-387, doi:10.1002/gepi.22048 (2017).
24 Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137, doi:10.1186/gb-2008-9-9-r137 (2008).
25 Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589, doi:10.1016/j.molcel.2010.05.004 (2010).
26 Yu, G., Wang, L. G. & He, Q. Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382-2383, doi:10.1093/bioinformatics/btv145 (2015).
27 Aue, A. *et al.* A Grainyhead-Like 2/Ovo-Like 2 Pathway Regulates Renal Epithelial Barrier Function and Lumen Expansion. *J Am Soc Nephrol* **26**, 2704-2715, doi:10.1681/ASN.2014080759 (2015).
28 Walentin, K. *et al.* A Grhl2-dependent gene network controls trophoblast branching morphogenesis. *Development* **142**, 1125-1136, doi:10.1242/dev.113829 (2015).
29 Gao, X. *et al.* Evidence for multiple roles for grainyhead-like 2 in the establishment and maintenance of human mucociliary airway epithelium. *Proceedings of the National Academy of Sciences* **110**, 9356-9361 (2013).
30 Senga, K., Mostov, K. E., Mitaka, T., Miyajima, A. & Tanimizu, N. Grainyhead-like 2 regulates epithelial morphogenesis by establishing functional tight junctions through the organization of a molecular network among claudin3, claudin4, and Rab25. *Mol Biol Cell* **23**, 2845-2855, doi:10.1091/mbc.E12-02-0097 (2012).
31 Varma, S. *et al.* The transcription factors Grainyhead-like 2 and NK2-homeobox 1 form a regulatory loop that coordinates lung epithelial cell morphogenesis and differentiation. *The Journal of biological chemistry* **287**, 37282-37295, doi:10.1074/jbc.M112.408401 (2012).
32 Cieply, B. *et al.* Suppression of the epithelial-mesenchymal transition by Grainyhead-like-2. *Cancer Res* **72**, 2440-2453, doi:10.1158/0008-5472.CAN-11-4038 (2012).
33 Xiang, X. *et al.* Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. *PLoS One* **7**, e50781, doi:10.1371/journal.pone.0050781 (2012).
34 Kumar, B. *et al.* Normal breast-derived epithelial cells with luminal and intrinsic subtype-enriched gene expression document inter-individual differences in their differentiation cascade. *Cancer Res*, doi:10.1158/0008-5472.CAN-18-0509 (2018).
35 Wang, X. *et al.* Epigenetic activation of HORMAD1 in basal-like breast cancer: role in Rucaparib sensitivity. *Oncotarget* **9**, 30115-30127, doi:10.18632/oncotarget.25728 (2018).
36 Kennecke, H. *et al.* Metastatic behavior of breast cancer subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **28**, 3271-3277, doi:10.1200/JCO.2009.25.9820 (2010).
37 de Silva Rudland, S. *et al.* Statistical association of basal cell keratins with metastasis-inducing proteins in a prognostically unfavorable group of sporadic breast cancers. *The American journal of pathology* **179**, 1061-1072, doi:10.1016/j.ajpath.2011.04.022 (2011).
38 Kao, J. *et al.* Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PLoS One* **4**, e6146, doi:10.1371/journal.pone.0006146 (2009).
39 Stingl, J., Eaves, C. J., Zandieh, I. & Emerman, J. T. Characterization of bipotent mammary epithelial progenitor cells in normal adult human breast tissue. *Breast Cancer Res Treat* **67**, 93-109 (2001).
40 Chung, V. Y. *et al.* The role of GRHL2 and epigenetic remodeling in epithelial-mesenchymal plasticity in ovarian cancer cells. *Commun Biol* **2**, 272, doi:10.1038/s42003-019-0506-3 (2019).

Supplemental data

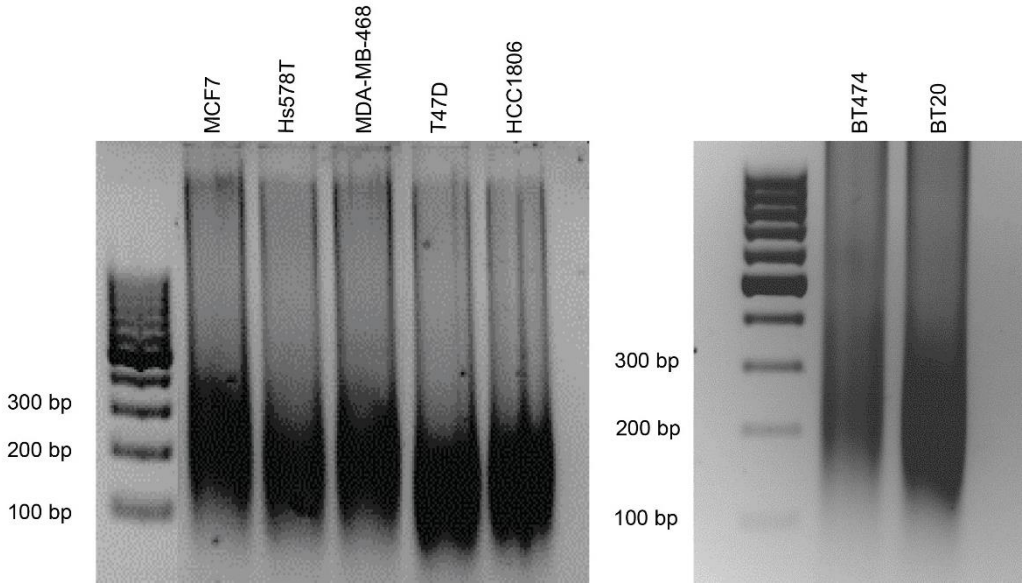
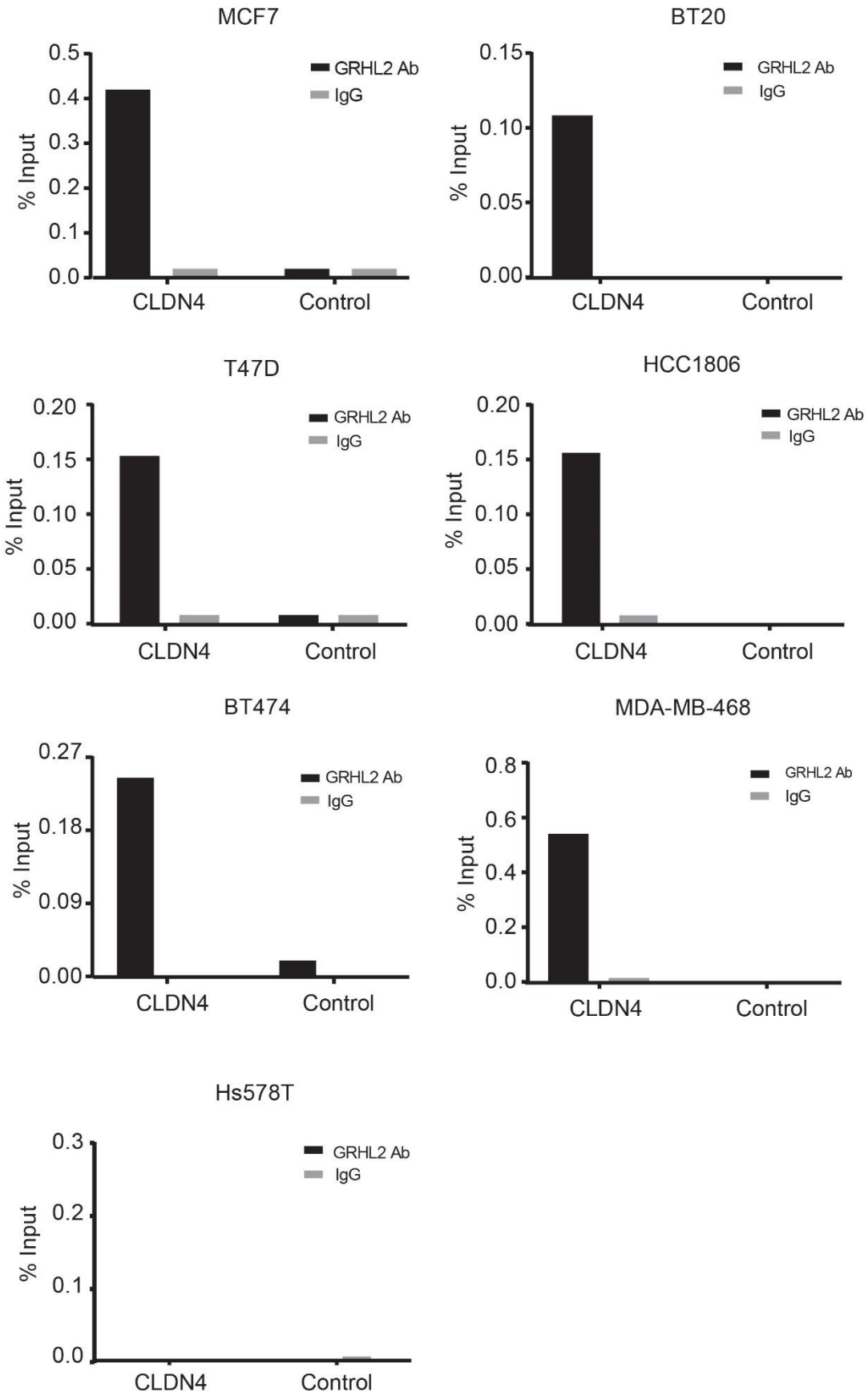


Fig. S1. DNA fragmentation analysis by agarose gel electrophoresis. After sonication, indicated samples were purified and loaded to 2% agarose gel.

(Next page)Fig. S2. ChIP-PCR validation of the isolated genomic DNA fragments. Graphs represent the efficiency of CLDN4 genomic DNA co-precipitation with anti-GRHL2 Ab (black bars) or IgG control Ab (grey bars). Detection was performed by PCR using primers targeting the promoter region of CLDN4 or targeting the intergenic region upstream of the GAPDH locus (Control). Results are shown for 3 GRHL2-positive luminal cell lines (MCF7, BT474, T47D), 3 GRHL2-positive basal-A cell lines (BT20, HCC1806, MDA-MB-468), and 1 GRHL2-negative basal-B cell line (Hs578T).

Fig. S2


MCF7
% Input
0.5
0.4
0.3
0.2
0.1
0.0
GRHL2 Ab
IgG
CLDN4
Control
BT20
% Input
0.15
0.10
0.05
0.00
GRHL2 Ab
IgG
CLDN4
Control
T47D
% Input
0.20
0.15
0.10
0.05
0.00
GRHL2 Ab
IgG
CLDN4
Control
HCC1806
% Input
0.20
0.15
0.10
0.05
0.00
GRHL2 Ab
IgG
CLDN4
Control
BT474
% Input
0.27
0.18
0.09
0.00
GRHL2 Ab
IgG
CLDN4
Control
MDA-MB-468
% Input
0.8
0.6
0.4
0.2
0.0
GRHL2 Ab
IgG
CLDN4
Control
Hs578T
% Input
0.3
0.2
0.1
0.0
GRHL2 Ab
IgG
CLDN4
Control

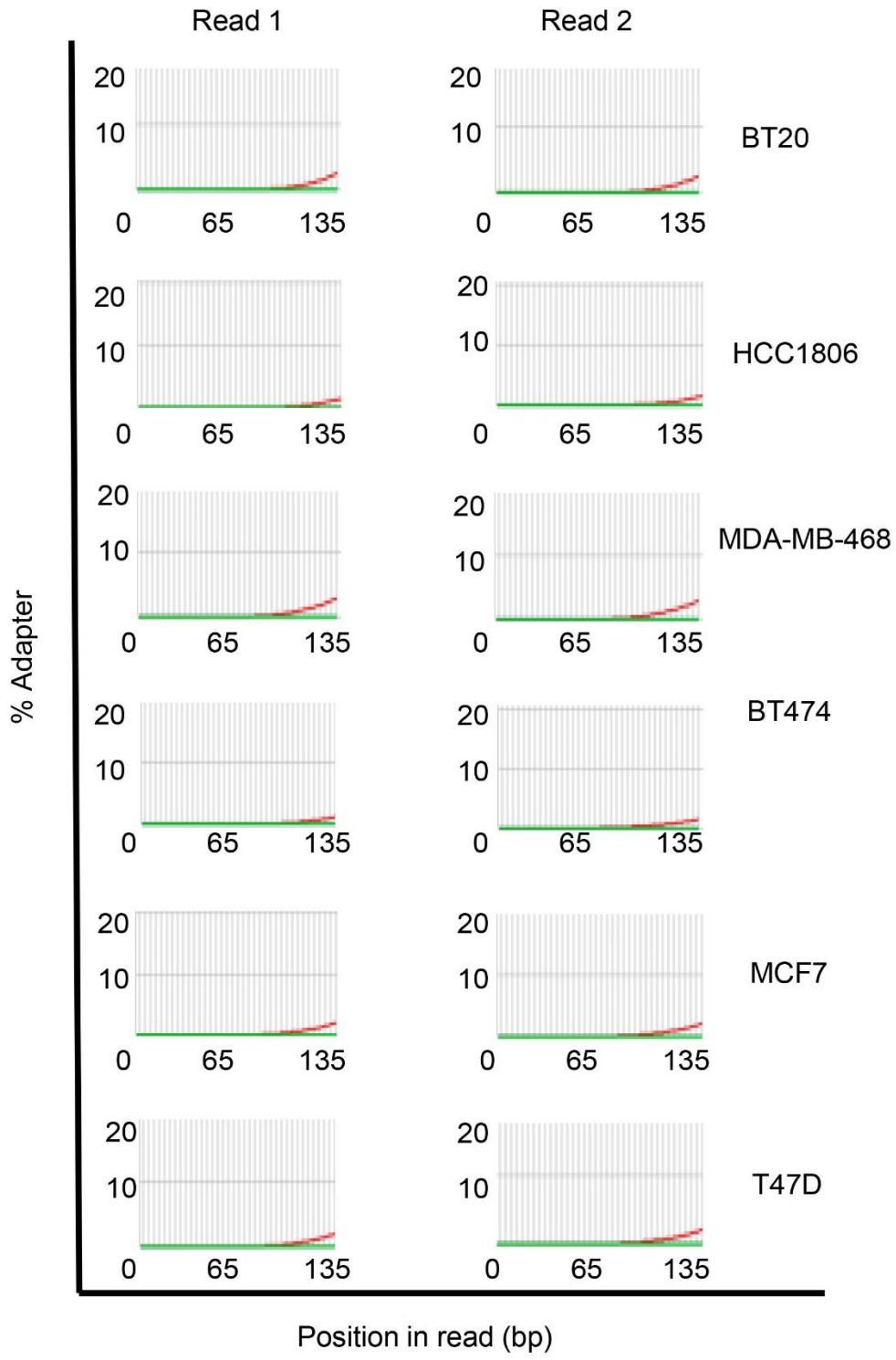


Fig. S3. Cumulative presence of adapter sequences. Results show that cumulative presence of adapter sequences is less than 5% in each cell sample, indicating that the data sets could be further analysed without adapter-trimming.

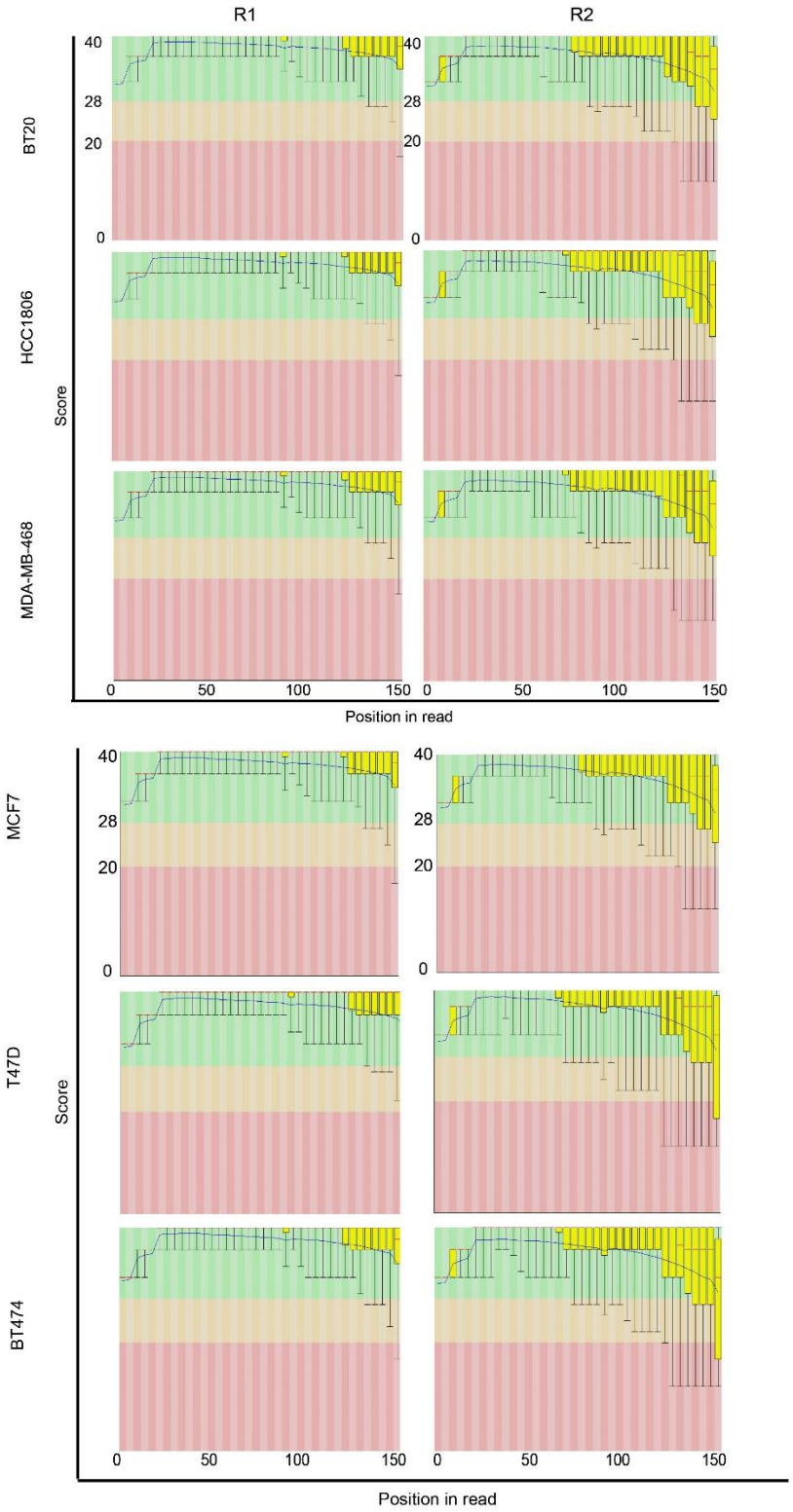


Fig. S4. Per base sequence quality for all sequencing data sets. Y axis is divided into good quality calls (coloured green), reasonable quality calls (coloured orange) and poor-quality calls (coloured red). In general, it is normal to observe base calls dropping into the red area towards the end of reads. The blue line representing the mean quality of base calls consistently stayed in the green area, indicating that sequencing data sets were of high quality.

Table S1. Candidate GRHL2 target genes. List of genes where GRHL2 was found to interact with the -1000 bp to +100 bp promoter region in all three basal A, all three luminal-like, or all basal A and luminal-like cell lines tested.

Specific target genes in basal A subtype:

TTK	RNF224	DHDDS	SUPT7L	TFDP2
CCDC171	RAB23	ILF2	NMRAL1	DNAH3
PATJ	ZNF436	SH2D4A	ZNF75D	CEP44
HRCT1	IVNS1ABP	ETV6	NFKBIZ	AP2S1
LOC100419583	SRFBP1	PLCD1	STXBP5-AS1	TFAP2A
PRG3	PKD1L2	NEK1	A2ML1	INHBA
CSTF2T	NCEH1	PYM1	TEX261	VAT1
NAPA	GGH	TMEM9	MUC20	HNRNPA1L2
TGM5	GOLGA5	ADGRL2	S100A9	MDGA1
HMGN4	PPP1R13L	WWP1	ATF3	MIR548XHG
LOC153684	ADGRF1	LOC283299	MTAP	LINC00989
KLHDC9	CDC25C	PDE12	GDF5	WDR5B
OTX1	FGD3	EYA2	LOC100128770	TSC22D1
LINC01637	SYTL2	SLPI	RAD23A	ZNF563
NUP155	LIMCH1	GTF3C2	DUSP10	PI3
ZNF700	PSMB6	CASC4	ADK	MINK1
MORN3	TMEM154	ADGRG7	PLS3	FER1L4
CEP162	ZNF468	TMEM173	RNF224	LOC101927822
CARD6	CLCN3	TENT4A	MCFD2	IGF2BP3
NXT1	DPY19L4	SPATA32	LINC01094	UBP1
C11orf52	LINC00474	KIAA0319	LINC01344	CBR4
ZNF280D	RNF19A	IMPDH1	ZNF682	SOX2
SEMA4B	DST	ALG10B	LEMD2	MIR5684
KIF16B	KLHDC7A	MORC4	THAP11	LOC105376114
ACSL1	TMEM256	HMGB2	PPP2R2C	SPINT2
MIR4432HG	TAB3	IFT22	GSDMC	SLC29A2
RPN2	OXR1	SYNE2	THADA	RGL1
RPSAP58	KMT5C	MICB-DT	SLC35F2	BCLAF1
COPS5	TMEM102	KHDC4	MIR7706	TOMM70
PUM2	SAMD12-AS1	ZNF221	PIGV	DHRS4-AS1
CFDP1	MUC15	BCAR4	ANAPC5	KLK8
SRP68	KIF5C	SCNM1	ETFBKMT	EPS8
PIGA	GLP2R	STK3	STK38L	POU2F3
RNVU1-15	GPBP1L1	EBPL	LOC93429	ZNF552
UFSP1	YWHAZ	ANKRD2	HRH1	ZNF562
HSPA4L	ACIN1	PTGR2	CRYM-AS1	ADGRE2
MIR3165	CITED4	FZR1	CAST	LOC103021296
ATP2C1	ADGRF2	MIR378J	TLCD2	VPS50
PGAM1P5	FLJ42969	SNX27	RXRA	SNX3
LYSMD3	Septin8	MUC1	ABT1	C11orf74
ZC3H15	FNTA	DTL	PHF23	LRG1
TRIT1	GS1-124K5.11	RPAIN	MDH1B	CEBPB-AS1
NFIA	HS3ST1	ABCA12	KRTDAP	KYAT3

ME1	IL36RN	LINC00869	B3GALT4	PHETA1
SLTM	ZNF695	MSANTD4	COL4A5	COTL1
RNF225	ZFAND6	LOC101928008	BTN3A3	PITPNM1
NFKBIA	BCAS2	MIR135B	PGRMC2	TMPRSS11E
PSORS1C1	GSE1	GLB1	ACKR2	PIP4P2
STAP2	RNU6-2	TGIF1	CD63	RASGEF1B
ZNF844	AFG3L2	ZNF284	PA2G4	CHRNE
DDX12P	UPK1A	GPR156	HMGCR	HS6ST2
BLNK	GTF2H4	WDR75	ZNF234	LDHA
MIR4422	ENO1-AS1	RCBTB1	TMPRSS4	RALY
RASL11A	ATG14	OSGIN1	ATAT1	PDE4DIP
MIR4799	IKBKE	HACE1	TRIM16L	LOC100506113
WBP1	BCL9	LINC01393	MTMR11	KIAA0513
UPK2	RBM47	PDSS1	LRRC23	PGLS
S100A12	ALKBH8	GCNT2	ANKMY1	INTS1
VOPP1	DTWD2	XDH	ZNF140	TM4SF4
ACE2	LARS	ATP1A1	RGS3	LINC01634
LOC101928977	EIF4G1	CRTAM	MON2	VSIG10L
UCA1	RNF222	CHD8	SLMO2-ATP5E	SLC2A11
IL1RN	EXOC6	TIGD2	LINC02447	TSHZ1
SCGB1B2P	MCCC1	ARHGAP27	ADAMTS6	GGTLC2
GJB3	SFTA2	IKBKG	EIF2AK2	KDM5C
SMIM13	DLGAP1-AS2	PDCD10	HIPK1-AS1	BNIPL
UNC13D	MGC32805	C7orf77	SEC11A	PEX26
MRPL1	RPIA	ELF3	DENND6A-DT	IMPDH1
FEZF1	CCDC26	GORAB	EEF2	KAT14
ALDH4A1	ACAD9	GTPBP3	PLS1	GBA
BEND5	ANXA11	PLEKHF2	LOC100294145	EIF1AD
RNF32	ANKRD54	S100P	CD164	LOC101927151
CSNK1G3	KDM6B	REXO2	LINC01588	MRPL24
CORO2B	RETREG1	SCAMP3	C1orf226	CAB39L
LINC01559	DEPDC1-AS1	CENPA	TCHHL1	IL17RE
LINC01354	SLC37A4	AATF	PPARD	TSKU

Specific target genes in luminal-like subtype

EPHA1	SSR4P1	OR7E91P	MIR6070	ARRDC3
FLJ31356	TMEM40	ADGRF4	TGM1	CLDN4
IQCK	ZNF440	MESP1	NIPSNAP1	MGP
LOC101927391	TGIF1	TUFT1	PTPN14	CARD14
NAALADL2	ZNF433	ARSD	FRRS1	CCDC12
RIMS1	SMG8	TIGAR	GAR1	ZNF823
EDEM2	PRR15L	TMEM79	PPOX	MIR4676
LINC01213	RBBP8NL	LOC101927318	TMPRSS11F	LINC00359
GRAMD2B	ANXA9	AMD1	MIR4513	PDGFB
NIPAL2	AFG1L	FAF1	ZNF799	LINC00456
SCAMP4	GGTLC1	AIFM1	IFRD1	ZNF274
VEPH1	DAZAP1	RPL41	ZNF20	TRPC4AP
KCNJ13	ZNF44	STX19	SYTL5	SLC4A7
ARHGAP32	NEU1	BATF	VGLL1	CMTR2

MIR6773	RPL32P3	ANKRD22	BCAS1	ZMYND8
TYSND1	OVOL2	LOC100506098	GPNMB	LRP10
CNP	LOC101927911	ELF5	CFAP45	EEA1
ADIPOQ	ST3GAL4	TMPRSS13	GPR108	LOC344967
LOC101927272	SLC9A1	RNVU1-14	NXT1	DLG4
BBOX1	SEMA4A	UBE2A	RAP2B	ERP27
CLDN8	ITFG2	IGSF9	EHF	EPN3
FMN1	CD46	ARHGEF19	LINC02408	PDCD2
PKP2	SLITRK6	ERBB3	PSCA	MAPK10
FKBP2	GRAMD1C	VIPAS39	EIF2B5	HRH1
LINC00346	DNAJC5B	DNAAF5	JUP	LIMA1
RBL2	ALDH3B2	ARHGAP24	EEF1E1	WSB2
EPB41L1	UBALD2	SBNO1	C1orf116	PGLYRP2
ZBTB20	LINC01405	GMPR2	TBL1X	YAP1
RASAL2	BMF	SNORA38	EHF	ROCK1P1
SLC25A45	P2RY6	SORT1	PLA2G4B	SLC41A3
ZER1	IKZF2	LOC100132781	RNU5B-1	ZNHIT6
ATAD3B	LINC00885	CHD3	LOC100129917	HIST2H2AB
GMEB1	C4orf3	IVL	RAB25	TRIL
CBLB	CDS1	MACROD1	KRT80	ZNF443
TJP2				

Common target genes between basal-like and luminal-like subtypes

ARHGEF38	ERLNC1	PIM2	LOC101927296	LOC102724064
GINS2	MIR6784	MTERF2	SLC40A1	PIK3C2G
PGR	NME7	CRISP3	DSCAM-AS1	LOC148709
MIR4328	KLK12	SLFN12	SFTPA2	LOC102724163
PROM2	ATP6V0A4	SLC10A5	TP53INP2	ZP1
MUCL1	PPEF1	FMO9P	PURG	ASCL2
DLX5	LINC00938	HIST2H2BF	PRIM2	JADE1
PDE4D	KRTAP3-1			