January 2015

Functional Analysis of the Ovarian Cancer Susceptibility Locus at 9p22.2 Reveals a Transcription Regulatory Network Mediated by BNC2 in Ovarian Cells

Melissa Buckley
University of South Florida, melissa.price22@yahoo.com

Functional Analysis of the Ovarian Cancer Susceptibility Locus at 9p22.2 Reveals a Transcription Regulatory Network Mediated by *BNC2* in Ovarian Cells

by

Melissa A. Buckley

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Cell Biology, Microbiology, and Molecular Biology
College of Arts and Sciences
University of South Florida

Major Professor: Alvaro N.A. Monteiro, Ph.D.
Committee: Jiandong Chen, Ph.D.
Javier Cuevas, Ph.D.
Cathy Phelan, Ph.D.

Date of Approval:
July 8, 2015

Keywords: Single Nucleotide Polymorphism, Allele,
Genome Wide Association Studies, Zinc Finger Domain

DEDICATION

I would like to dedicate my thesis to my friends and family who have supported me throughout the whole process. First and foremost a very special thanks to my husband, Kip. It takes a lot of perseverance and patience to be the spouse of a PhD candidate and I am so thankful I have married the most patient person I know. You have kept me grounded throughout the process and made me laugh and ensured me that we would get through it. I also want to thank my daughter for emulating me and therefore motivating me to reach for the best and be my best self. I also want to thank my mom and dad, Harry and Kelly, for their enormous support throughout my PhD by providing a home away from home for my husband working in another city and helping take care of my daughter while I learned to become Mommy, PhD. I also want to thank my parents and grandparents and brothers for all their love and encouragement throughout my life in which without it, I would not have made it this far. I also want to thank my mother and father in law for their love, support, and encouragement. You truly have treated me like another daughter in giving me advice and reminding me how proud you are of me.

I also want to thank my good friend, Christine, whom in 3rd grade started to motivate me on this path. We were after school science lab partners in elementary school to room-mates and best friends in college studying pre-med. Thank you for being there for me and having fun with me in our free time. Last but not least, I would like to thank my best friend Jessica. They say people develop bonds when under stressful or difficult circumstances. I think grad school must fall under this category since we truly developed a strong friendship from the very beginning of our time at Moffitt. Thanks for being a shoulder to cry on when experiments failed again and again. I am glad I had someone to run marathons with to de-stress and I am so glad we got to share experiences of getting married, having children, and getting those three letters added to the end of our name.

ACKNOWLEDGMENTS

First I would like to thank my mentor, Dr. Monteiro for guiding me through this process, sharing all his knowledge and stories, and providing a tremendous amount of support through tough situations. I would also like to thank my committee members, Jiandong Chen, Javier Cuevas, and Catherine Phelan for their guidance. I would also like to thank the Monteiro lab members, present and past, Rebekah Baskin, Olga Golubeva, Ankita Jhuraney, Alex Valle, Renato Carvalho, and Aneliya Velkova with a special thanks to Gustavo Mendoza-Fandino, Anxhela Gjyshi, Xueli Li, and Paulo Cilas M. Lyra for helping perform difficult experiments that contributed to this work. I am glad we all had such a great sense of comradery in the lab which made graduate school much more pleasant. I would also like to thank a former post-doc in the lab, Nicholas Woods, and a former post-doc in a collaborating lab Howie Shen for being my second mentors. They taught me how to do experiments, provided insight and ideas, and also advised how to maintain a work-life balance in academia. Additionally, thank you Nick for helping the project move off the ground. I am pretty sure it would still be there if it was not for you.

Additional experimental support has been provided by Kate Lawrenson, Dennis Hazelett, Simon G. Coetzee, and Houtan Noushmehr. They provided the RNA-seq, FAIRE-seq, and ChIP-seq for histone markers data to support and base my scientific hypotheses on. Ally Yang and Hamed Shateri Najafabadi from Timothy Hughes lab also

performed the cutting edge experiments with the PBM and bioinformatics analysis of the zinc finger binding sites.

I would also like to thank members of the Ovarian Cancer Association Consortium, specifically, Thomas Sellers, Catherine Phelan, Simon Gayther, Paul Pharoah, and Ellen Goode for their guidance and for allowing me to contribute to the cause of the consortium. It has been quite the learning experience which I do not think many graduate students get to encounter. It has been really special to see the small contribution I am making, influences a greater cause. I would also like to thank the NIH for financial support with the Ruth L. Kirschstein National Research Service Award. Also, thank you ARCS (Achievement Rewards for College Scientists) Foundation for their financial support and striving to promote science in the United States.

TABLE OF CONTENTS

TABLE OF CONTENTS i

LIST OF TABLES iv

LIST OF FIGURES v

ABSTRACT viii

CHAPTER ONE: BACKGROUND 1
- Ovarian Cancer 1
  - Ovarian Cancer Subtypes 1
  - Risk Factors and Pathogenesis 3
  - Genetic Predisposition to Ovarian Cancer 5
- Genome Wide Association Studies 7
  - Linkage Analysis versus GWAS 8
  - Principles of GWAS 10
  - Caveats of GWAS 12
  - Ovarian Cancer GWAS 13
  - Functional Analysis of Susceptibility Loci 16
- Transcriptional Regulation 18
  - Basics of Transcriptional Regulation 18
  - Transcription Factor Binding 21
  - Chromatin Structure 22
  - 3D Structure of the Genome 24

CHAPTER TWO: IDENTIFYING FUNCTIONAL SNPS 27
- Introduction 27
- Results 28
  - Candidate Functional SNPs 28
  - Functional Analysis of SNPs 31
    - Mapping SNPs to regulatory elements 31
    - Development of Enhancer Scanning method 33
      - Experimental Design 33
      - Choice of template, tile verification, and host cell 35
      - Controls 36
      - Statistical analysis 37
      - Anticipated results 37
      - Limitations 38
    - Mapping SNPs to regions of enhancer activity 40
  - SNPs with Allele Specific Effects 40

SNPs with allele specific enhancer activity 40
Allele specific activities in electrophoretic mobility shift assays 42
Summary 42
Materials and Methods 45
Candidate Functional SNPs 45
Cell Lines and Cell Type-specific Datasets 45
FAIRE-Seq and ChIP-Seq for Histone Modifications 46
Enhancer Scanning 46
Genome Browser 48
Electrophoretic Mobility Shift Assays 48

CHAPTER THREE: IDENTIFICATION OF TARGET GENE 50
Introduction 50
Results 51
BNC2 and CNTLN as Candidate Target Genes 51
Transcription activity at candidate gene promoters 51
Expression Quantitative Trait Loci (eQTL). 51
Region 8 is in Physical Proximity to the TSS of *BNC2* in Ovarian Cells 58
Summary 54
Materials and Methods 55
eQTL Analysis 55
3C 55

CHAPTER FOUR: FUNCTIONAL ANALYSIS OF BNC2 57
Review of *Basonculin 2* 57
Identification of *Basonuclin 2* and comparison to *Basonculin 1* 57
Expression of *BNC2* 59
Function of BNC2 60
Results 63
C2H2 Zinc Finger Proteins 63
BNC2 Zinc Fingers Recognize Specific DNA Sequences *in vitro* 65
BNC2 Genome-wide Target Sites 68
Identification and Validation of BNC2 Target Genes 70
BNC2 Interacts with the NuRD Complex 75
Summary 77
Materials and Methods 78
Protein Binding Microarray 78
ChIP/ChIP-Seq for BNC2 79
Nanostring 82
Tandem Affinity Purification coupled to LC-MS/MS 83
Transcriptional Repression Assay 85

CHAPTER FIVE:DISCUSSION 90

REFERENCES... 96

ABOUT THE AUTHOR...END PAGE

LIST OF TABLES

Table 1: Candidate Functional SNPs ... 28

Table 2: Twenty-two SNPs associated with ovarian cancer risk overlap with areas of regulatory activity ... 32

Table 3: Expression analysis of downstream target genes of BNC2 ... 72

Table 4: Genes with significant expression changes upon differential expression of *BNC2* ... 76

Table 5: List of oligos used in this study ... 85

LIST OF FIGURES

Figure 1: Effect size versus allele frequency of ovarian cancer susceptibility genes 6

Figure 2: Penetrance of inherited variants........ 8

Figure 3: Tagging SNP represents many candidate causal SNPs in LD........ 11

Figure 4: EOC Susceptibility Loci........ 14

Figure 5: Flow chart for the functional analysis of ovarian cancer susceptibility loci 17

Figure 6: Regulatory regions and how they are identified 24

Figure 7: Luciferase assay protocol........ 35

Figure 8: Functional annotation identifies candidate functional SNPs overlapping with regions of regulatory activity in ovarian cells........ 39

Figure 9: SNPs showing allele-specific activities........ 41

Figure 10: Conservation and S/MAR predicted sequences within the locus 43

Figure 11: Summary of location of SNPs with most compelling evidence for function........ 44

Figure 12: Functional SNPs influence transcription of *BNC2* 52

Figure 13: Summary of locus target genes........ 54

Figure 14: BNC2 binds to its own promoter........ 66

Figure 15: BNC2 recognizes specific nucleotide sequence........ 69

Figure 16: Kegg Pathway analysis of downstream targets of BNC2 71

Figure 17: Overexpression of *BNC2* in HEK293FT cells........ 72

Figure 18: BNC2 interacts with the NuRD complex........ 77

Figure 19: Outline of the study 95

ABSTRACT

GWAS have identified several chromosomal loci associated with ovarian cancer risk. However, the mechanism underlying these associations remains elusive. We identify candidate functional Single Nucleotide Polymorphisms (SNPs) at the 9p22.2 ovarian cancer susceptibility locus, several of which map to transcriptional regulatory elements active in ovarian cells identified by FAIRE-seq (Formaldehyde assisted isolation of regulatory elements followed by sequencing) and ChIP-seq (Chromatin Immunoprecipitation followed by sequencing) in relevant cell types. Reporter and electrophoretic mobility shift assays (EMSA) determined the extent to which candidate SNPs had allele specific effects. Chromosome conformation capture (3C) reveals a physical association between *Basonuclin 2 (BNC2)* and SNPs with functional properties. This establishes *BNC2* as a major target of four candidate functional SNPs in at least two distinct elements.

BNC2 codes for a putative transcription regulator containing three pairs of zinc finger (ZF) domains. Furthermore, *bnc2* mutation in zebrafish leads to developmental defects including dysmorphic ovaries and sterility, clearly implicating this protein in cellular processes associated with ovarian development. We show that BNC2 is a transcriptional regulator with a specific DNA recognition sequence of targets enriched in genes involved in cell communication through DNA binding assays, ChIP-seq, and expression analysis.

This study reveals a comprehensive regulatory landscape at the 9p22.2 locus and indicates that a likely mechanism of susceptibility to ovarian cancer may include multiple allele-specific changes in DNA regulatory elements some of which alter *BNC2* expression. This study begins to identify the underlying mechanisms of the 9p22.2 locus association with ovarian cancer and aims to provide data to support advances in care based on one’s genetic composition.

CHAPTER ONE:

BACKGROUND

Ovarian Cancer

Ovarian cancer is one of the leading causes of cancer deaths among women in the United States. It is a poorly understood disease often diagnosed at late stages and consequently with low five year survival rates (Vaughan et al., 2011). In fact, the five year survival rate has plateaued at 40% since the introduction of platinum based therapies in the late 1970s (Vaughan et al., 2011). One reason for the low survival and lack of improvements in therapy is that ovarian cancer has been treated as one disease when in actuality it is made up of different subtypes with diverse cellular origins and molecular pathways altered (Berns and Bowtell, 2012; TCGA, 2011). Additionally the pathogenesis of ovarian cancer is unclear. Identifying risk factors and those with a genetic predisposition would aid in identifying the disease and therefore lead to increased survival.

Ovarian Cancer Subtypes

Four different subtypes, mucinous, clear cell, endometrioid, and serous pertain to ovarian cancer. Evidence suggests that the mucinous subtype derives from metastases to the ovary from gastrointestinal tumors (Kelemen and Kobel, 2011; Lee and Young, 2003). Clear cell and endometrioid subtypes originate in the endometrium and are often linked with endometriosis (Nezhat et al., 2008). The most common and most lethal subtype, high grade serous ovarian cancer, was initially thought to originate in the ovarian surface epithelium (OSE). The cancer is often found at late stages and fills the peritoneal cavity making it difficult to discern the cell of origin. Yet, significant evidence also supports that the secretory cells in the fimbria of the fallopian tube contribute to the origin of high grade serous ovarian cancer (Carlson et al., 2008; Lee et al., 2007; Piek et al., 2001).

High grade serous ovarian cancer in itself seems to be very heterogeneous in that many patients experience different outcomes (Tan et al., 2013). Indeed, high grade serous ovarian cancer is made up of different molecular subtypes defined by expression analysis and clustering of ovarian tumor samples (Leong et al., 2015; Tan et al., 2013; TCGA, 2011; Tothill et al., 2008). These molecular subtypes include mesenchymal, immuno-reactive, differentiated, and proliferative (Leong et al., 2015; TCGA, 2011). The mesenchymal subtype displays severe myofibroblast infiltration and has an epithelial to mesenchymal gene expression signature (Leong et al., 2015; Tan et al., 2013; TCGA, 2011; Tothill et al., 2008). The WNT/beta-Catenin and Extra Cellular Matrix Pathways are altered and *HOX* genes and *FAP*, a stromal component, are over-expressed (TCGA, 2011; Tothill et al., 2008). This subtype also has poor prognosis (Leong et al., 2015; Tan et al., 2013; Tothill et al., 2008). The immune-reactive subtype displays T-Cell infiltration and expresses T-Cell chemokine ligands (Leong et al., 2015; TCGA, 2011; Tothill et al., 2008). This subtype has an intermediate prognosis (Leong et al., 2015; Tothill et al., 2008). The differentiated subtype expresses ovarian tumor markers *MUC1* and *MUC16* as well as the fallopian tube marker *SLP1* (Leong et al., 2015;

TCGA, 2011). This subtype also has an intermediate response (Leong et al., 2015). The last subtype, proliferative, expresses stem cell factors including transcription factors (TFs) *HMGA2* and *SOX11* (Leong et al., 2015; TCGA, 2011) This subtype also has low expression of ovarian tumor markers *MUC1* and *MUC16* and high expression of proliferation markers *MCM2* and *PCNA* (TCGA, 2011). This subtype has a poor prognosis yet seems to be sensitive and respond well to vinca alkaloids (Leong et al., 2015; Tan et al., 2013).

Risk Factors and Pathogenesis

Since high grade serous epithelial ovarian cancer (EOC) is diagnosed at an advanced stage in 70% of patients and these patients have a worse outcome than those with early stage disease (Vaughan et al., 2011), identifying those at risk may lead to early detection and therefore decreased mortality. One known risk factor that significantly influences ovarian cancer is low parity (Braem et al., 2010; Hinkula et al., 2006; Salehi et al., 2008; Sueblinvong and Carney, 2009). In fact, women who have children have a decreased risk of 71% and the risk further decreases by 10% with each live birth (Braem et al., 2010). Oral contraceptive use and shorter menstrual lifespan also decrease risk of ovarian cancer (Bosetti et al., 2002; Braem et al., 2010; Hankinson et al., 1992; Modugno et al., 2004; Salehi et al., 2008; Sueblinvong and Carney, 2009). Pregnancy, oral contraceptives, and shorter menstrual lifespan all decrease the number of ovulations in a lifetime.

There is no clear evidence for the pathogenesis of ovarian cancer but several hypotheses exist based on the number of ovulations/menstrual cycles in a lifetime. The first hypothesis states that incessant ovulation leads to damage of the ovarian surface epithelium which in turn leads to malignant cells (Fathalla, 1971; Riman et al., 1998). Another hypothesis suggests that granulosa and theca cells fail to undergo apoptosis after ovulation (Cramer et al., 2002; Hanna and Adams, 2006). Another hypothesis suggests that high levels of gonadotropins increase stimulation of estrogen, which entraps ovarian epithelial cells in inclusion cysts which leads to malignant cells (Hanna and Adams, 2006; Zheng et al., 2007). Higher androgen levels lead to cancer while higher levels of progestin prevent cancer (Bu et al., 1997; Hanna and Adams, 2006; Risch, 1998; Zheng et al., 2007). Pregnancy and oral contraceptives decreases gonadotropin and androgen levels while increasing levels of progestin, therefore preventing cancer (Sueblinvong and Carney, 2009). Finally, inflammation potentially plays a major role in ovarian cancer development (Hanna and Adams, 2006; Ness et al., 2000; Salvador et al., 2009). During menstruation, retrograde flow brings inflammatory mediators (bacteria, chemicals, etc.) to the fallopian tube and therefore inflammation within the fallopian tube (Maisey et al., 2003; McGee et al., 1999; Salvador et al., 2009; Strandell et al., 2004). Inflammation within the fallopian tube causes cells to rapidly divide and increase the potential for DNA replication errors and thus development of a malignant cell (Ames et al., 1995; Dreher and Junod, 1996; Nash et al., 1999; Pagano et al., 2004; Salvador et al., 2009). Oral contraceptives and pregnancy decrease and eliminate menstruation respectively and therefore decrease retrograde flow (Brosens and Vasquez, 1976; Group, 2005; Lindblom et al., 1980; Salvador et al., 2009). Also oral contraceptives and pregnancy increase the cervical mucus thickness which protects the uterus from inflammatory mediators (Pal and

Bhattacharyya, 1989; Salvador et al., 2009). This theory has the most effect on the fallopian tube suggesting that inflammation may be the most likely pathogen for ovarian cancer (Salvador et al., 2009).

Genetic Predisposition to Ovarian Cancer

Family history is also a risk factor for ovarian cancer. Ovarian cancer is a seemingly inherited disease in that women with a first degree affected relative have an increased risk of developing ovarian cancer compared to the general population (Pharoah and Ponder, 2002; Stratton et al., 1998). Risk further increases in families with mutations in *BRCA1* and *BRCA2* by 45% and 25% respectively (Antoniou et al., 2002; Ford et al., 1998; Minion et al., 2015; Pharoah and Ponder, 2002). Germline mutations in mismatch repair genes (*MMR*) genes such as *MLH1*, *MSH2*, *MSH6*, *PMS2*, and *EPCAM*, also known as Lynch Syndrome causes 2% of ovarian cancer cases (Lu and Daniels, 2013; Malander et al., 2006; Minion et al., 2015; Pennington and Swisher, 2012; Pennington et al., 2014; Walsh et al., 2011). Lynch syndrome was first identified in families with colorectal cancer and the ovarian cancer patients with mutations in *MMR* genes have a family history of colorectal cancer (Malander et al., 2006; Meyer et al., 2009). Mutations in *BRCA1* and *BRCA2* lead to high grade serous ovarian cancer while mutations in *MMR* genes more likely lead to endometrioid and mucinous ovarian cancer (Berns and Bowtell, 2012; Chiaravalli et al., 2001; Fujita et al., 1995; King et al., 1995; Turner et al., 2004).

Highly penetrant pathogenic alleles of known susceptibility genes such as *BRCA1*/*BRCA2* and *MMR* genes only account for 11% and 2% of high grade serous

EOC and endometrioid/mucinous in the general population, and less than half of all familial ovarian cancer cases (Malander et al., 2006; Pharoah and Ponder, 2002; Ramus et al., 2007). These genes have been identified by performing family-based linkage studies (Miki et al., 1994; Wooster et al., 1995). Germline mutations in highly

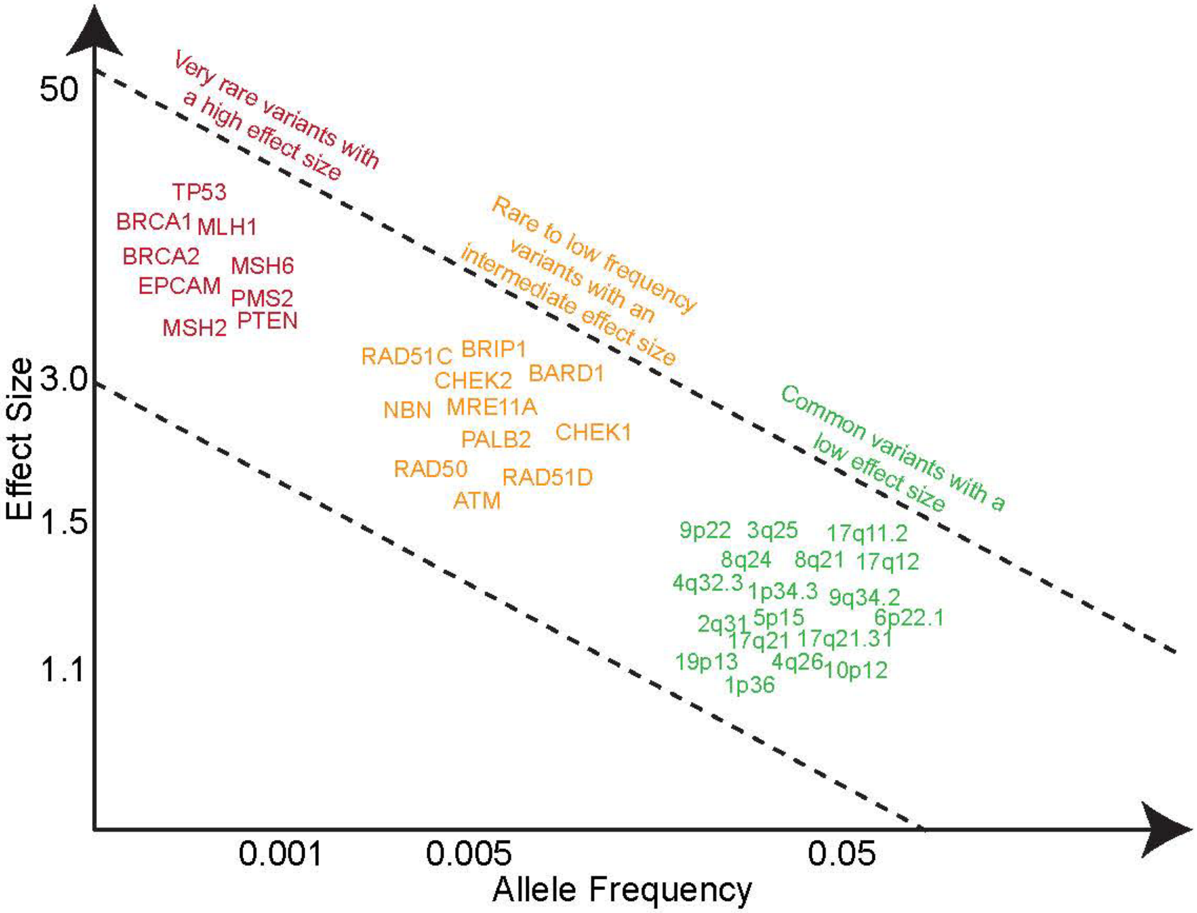


Figure 1. Effect size versus allele frequency of ovarian cancer susceptibility genes. This figure displays all known ovarian cancer susceptibility genes graphed by their approximate effect size and allele frequency in the population. This mainly portrays that the genetic contribution to risk includes high effect size/penetrance genes, intermediate effect size/penetrance genes and low effect size/penetrance loci.

penetrant cancer susceptibility genes *TP53* and *PTEN* have also been found in ovarian cancer cases (Minion et al., 2015; Pennington and Swisher, 2012; Pennington et al., 2014). However, exhaustive family-based linkage studies have not found novel highly penetrant genes (Pharoah et al., 2004). Additional studies have found variants with intermediate penetrance affecting ovarian cancer risk including several genes in the Fanconia Amenia pathway; *BARD1*, *BRIP1*, *MRE11A*, *NBN*, *PALB2*, *RAD50, RAD51C*, *RAD51D*, *CHEK2*, as well as *CHEK1* and *ATM* (Baysal et al., 2004; Casadei et al., 2011; Castera et al., 2014; Coulet et al., 2013; Kanchi et al., 2014; Kuusisto et al., 2011; Meindl et al., 2010; Pennington and Swisher, 2012; Pennington et al., 2014; Rafnar et al., 2011; Thorstenson et al., 2003; Walsh et al., 2011) Mutations in genes like *BRCA1* and *BRCA2,* with a high effect size or high penetrance, are very rare in the population (Figure 1) (Manolio et al., 2009). Mutations in intermediate penetrance genes are usually rare to low frequency variants (Figure 1) (Manolio et al., 2009). The remaining genetic contribution to risk in ovarian cancer may be explained by common variants with a low effect size (Figure 1) (Manolio et al., 2009).

Genome Wide Association Studies

Genome wide association studies (GWAS) identify the common variants associated with complex, common disease, such as ovarian cancer (Cardon and Bell, 2001; Pharoah et al., 2004; Risch and Merikangas, 1996; Risch, 2000). GWAS identify predisposition loci by genotyping thousands of SNPs in thousands of cases and thousands of controls to find the common low penetrant alleles that significantly occur more frequently in the cases than in the controls (Cardon and Bell, 2001; Carlson et al., 2004; Pharoah et al., 2004; Risch and Merikangas, 1996; Risch, 2000).

Linkage Analysis versus GWAS

Linkage analysis identifies genetic variants associated with diseases that follow the mendelian pattern of inheritance, meaning one gene affects one trait with two (or

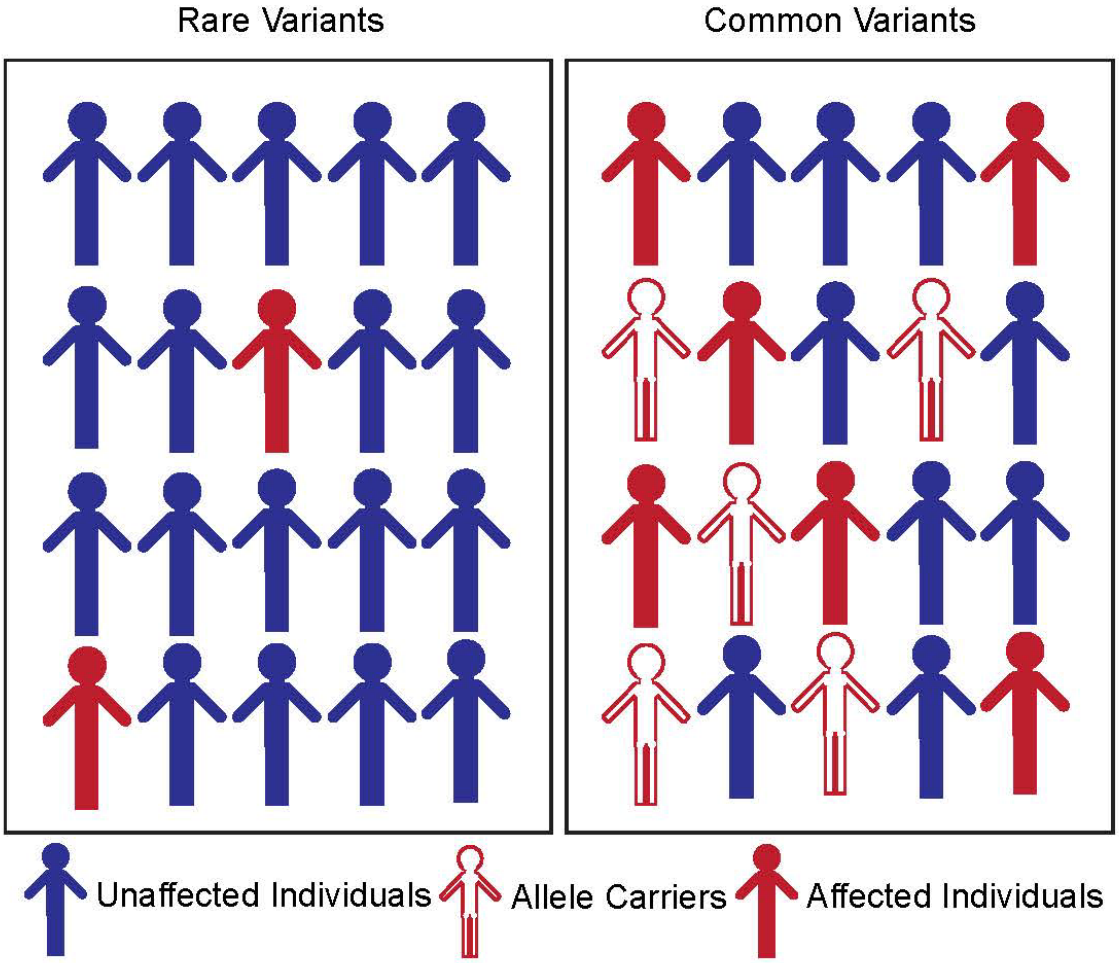


Figure 2. Penetrance of inherited variants. Carriers of variants (individuals outlined in red) can make up a small proportion of the population or larger proportion of the population depending on the type of variant and how it influences disease and fitness. Rare variants tend to have high penetrance with the majority of carriers developing the disease (red individuals) while common variants tend to have low penetrance.

very few) phenotypes (Jimenez-Sanchez et al., 2001). The disease associated variant is rare in the population and highly penetrant or in other words, carriers of the variant have a high likelihood of developing the disease (Pritchard, 2001; Reich and Lander, 2001) (Figure 2). Linkage analysis utilizes pedigrees of affected families to identify genetic markers with a known location in the genome that inherit with the disease gene during meiosis. Since regions of the genome that are in close proximity are less likely to recombine onto separate chromosomes, this allows for the identification of a location for the disease gene. Sequencing and identification of variants in the human genome attributed to the success of linkage analysis (International HapMap, 2005; Lander et al., 2001; Venter et al., 2001). Variations identified via linkage analysis often change protein coding sequences and therefore affect the function of the gene.

GWAS analysis identifies genetic variants associated with complex, common diseases or diseases that follow a non-Mendelian pattern of inheritance (Cardon and Bell, 2001; Manolio, 2010; Pharoah et al., 2004). These traits are polygenic, meaning multiple genes affect one trait with a wide range of phenotypes. The disease associated variants are common in the population, yet, a small proportion of carriers of a single disease associated variant will develop the disease since these variants, alone, have a small effect size (Cardon and Bell, 2001; Risch and Merikangas, 1996) (Figure 1 and 2). In order to obtain the statistical power that a variant associates more frequently with the disease than controls, GWAS genotype thousands of SNPs (common SNPs with a minor allele (MAF) > 0.05) in thousands of cases and thousands of controls (Cardon and Bell, 2001; Carlson et al., 2004; Colhoun et al., 2003; Dahlman et al., 2002; Hunter and Kraft, 2007; Manolio, 2010; Pharoah et al., 2004; Risch and Merikangas, 1996).

Statistical power also depends on the effect size (Stranger et al., 2011). By increasing the number of cases and controls, a GWAS will identify variants the have a smaller effect size (Manolio, 2010). In GWAS, the causal SNPs often change the sequence of non-coding DNA (Hardy and Singleton, 2009; Manolio, 2010). In general changing the sequence of non-coding DNA, has a less dramatic effect than changing an amino acid sequence of a protein, thus differences in effect size.

Principles of GWAS

The variant identified as associated with the disease in GWAS is called the tagging SNP. The tagging SNP represents all linked SNPs or in other words all SNPs that frequently inherit with that SNP after recombination during meiosis (Carlson et al., 2004; Gabriel et al., 2002). Regions of the genome being inherited together rather than independently is a phenomenon called linkage disequilibrium (LD). The tagging SNP tells us that any variation within a LD structure could be considered the causal variant (Carlson et al., 2004; Gabriel et al., 2002; Pharoah et al., 2004) (Figure 3). This allows for the genotyping of a subset of SNPs rather than all SNPs within the genome (Carlson et al., 2004; Pharoah et al., 2004). The HapMap project determined patterns of LD in the human genome among different ethnicities to allow for the selection of SNPs in GWAS (International HapMap, 2005). When selecting the tagging SNPs, it is important to keep in mind that patterns of LD are different among ethnicities (Carlson et al., 2004).

R^2 and d-prime are pairwise measurement of LD between SNPs. If $r^2 = 1$ and d-prime = 1 then two of four possible haplotypes are present (Pharoah et al., 2004). The first haplotype would be the major allele for both SNPs and the second haplotype would

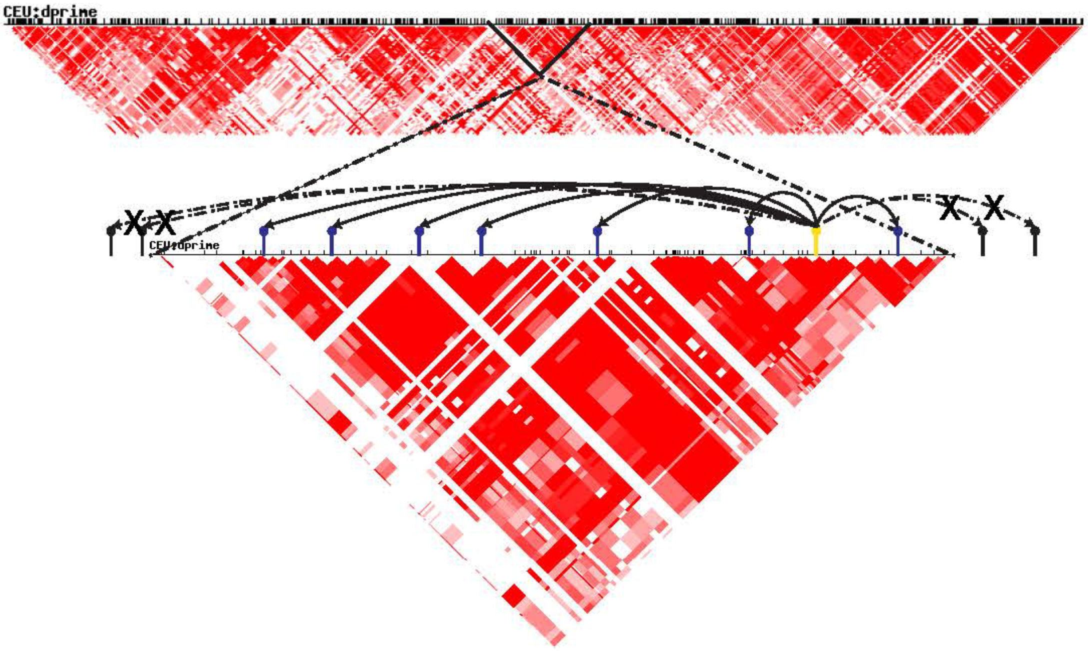


Figure 3. Tagging SNP represents many candidate causal SNPs in LD. SNPs that frequently inherit together during meiosis are in LD and are represented by LD structures. The increase in red intensity in the LD plot shown here, indicate SNPs in increasing LD measured by d-prime. One can visualize LD structure by breaks in red color. A tagging SNP (in yellow) would represent all SNPs within the LD structure (in blue) as potential candidate causal SNPs. SNPs outside that LD structure (in black) are not potentially causal.

be the minor allele for both SNPs. Neither site has experienced mutation or recombination between the sites (Carlson et al., 2004). If d-prime and $r^2 < 1$ then more than two haplotypes are present (Pharoah et al., 2004). Allele frequency also influences r^2 since a low frequency allele is less likely to occur in a haplotype (Pharoah et al., 2004). $r^2 = 1$ only when the two variants have the same MAF (Carlson et al., 2004).

To manage the high cost of such studies, the analysis is often done in tiers (Manolio, 2010; Pharoah et al., 2004). The first tier is testing all possible tagging SNPs throughout the genome in a smaller set of cases and controls. Those SNPs that surface as significant at an arbitrary threshold are genotyped again in a larger set of cases and controls. The final tier genotypes the SNPs that remained significant at a higher arbitrary threshold in tier 2 in an even larger set of cases and controls with the final SNPs reaching genome wide significance (Manolio, 2010; Pharoah et al., 2004). This process eliminates false positives and allows false negatives to surface (Manolio, 2010). Additionally, meta-analyses of independent GWAS increase sample size and statistical power (Stranger et al., 2011).

SNPs with a p-value < 5 x 10^{-8} are considered genome-wide significant. This is a Bonferroni correction based on multiple testing hypothesis since a GWAS genotypes approximately 1 million SNPs which, based on LD, seems to provide complete coverage of the genome (Hirschhorn and Daly, 2005; Stranger et al., 2011).

Caveats of GWAS

After hundreds of GWAS, many diseases still have missing heritability due to the lack of statistical power of a causal allele. The causal allele may not have been thoroughly represented by a tagging SNP due to correlation and/or differences in allele frequency (Stranger et al., 2011). One possibility for missing heritability is that there are rare variants with modest effect which would not be detected by GWAS (Manolio et al., 2009). The 1000 Genomes Project has sequenced 2,000 individuals to begin identifying low frequency variants (MAF = 0.001 – 0.005) but detecting association would still require sequencing of thousands of cases and controls (Genomes Project et al., 2010; Manolio et al., 2009). GWAS also lacks the ability to detect structural variants or copy number variation (Manolio et al., 2009). It is still unclear whether multiple variants associated with the disease are additive or non-additive in measuring susceptibility.

Additive would mean that each variant has an effect size and carriers of more than one disease variant would have a combined effect size of those variants (Hirschhorn and Daly, 2005; Manolio et al., 2009; Stranger et al., 2011). If some variants are non-additive, determining risk becomes more complex. Non-additive genetic variance includes interactions between loci such as dominance and epistasis as well as interaction between loci and environment (Hirschhorn and Daly, 2005; Manolio et al., 2009; Stranger et al., 2011). If the latter is the case, then fewer variants are required to explain the heritability (Hirschhorn and Daly, 2005). If multiple alleles at a disease susceptibility locus confer susceptibility, the power of LD mapping approaches decrease because each variation will arise on a different haplotype background (Pharoah et al., 2004). Since for many diseases there is still a gap in the heritability, risk assessment for individuals will not be of clinically useful predictive value (Jakobsdottir et al., 2009; Manolio, 2010; Wray et al., 2008).

Ovarian Cancer GWAS

Genome-wide association studies (GWAS) have identified a total of 20 loci associated with risk of ovarian cancer (Figure 1 and 4) (Bojesen et al., 2013; Bolton et al., 2010; Chen et al., 2014; Kuchenbaecker et al., 2015; Permuth-Wey et al., 2013; Pharoah et al., 2013; Shen et al., 2013; Song et al., 2009). The first ovarian cancer GWAS was done in three tiers and identified the locus 9p22 as associated with ovarian cancer (Song et al., 2009). They then re-analyzed the data to see if the associated SNP was more associated with a specific subtype of ovarian cancer (Song et al., 2009). It was only significantly associated with the serous ovarian cancer subtype (Song et al., 2009).

Another GWAS that looked at survival in ovarian cancer identified a SNP on 19p13 (Bolton et al., 2010). Yet, this SNP did not replicate in the final study (Bolton et al., 2010). It did replicate as associated with serous ovarian cancer (Bolton et al., 2010).

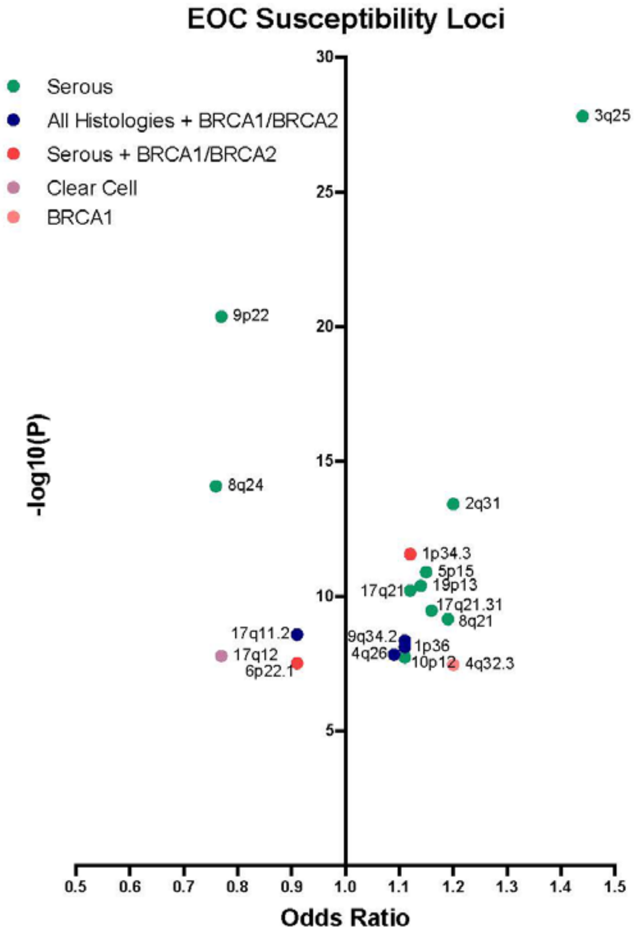


Figure 4. EOC Susceptibility Loci. Each EOC susceptibility loci identified in ovarian cancer GWAS has been plotted here based on their odds ratio (frequency in cases/frequency in controls) and EOC GWAS significance. These loci are associated with serous EOC, *BRCA1*/*BRCA2* mutation carriers and all histoligies, *BRCA1*/*BRCA2* mutation carriers and serous EOC, clear cell EOC, and *BRCA1* mutation carriers.

After these initial ovarian cancer GWAS, investigators realized that stratifying by subtype may allow for the discovery of more associated SNPs since phenotypic heterogeneity can reduce power (Goode et al., 2010; Ioannidis et al., 2009). Many of these loci also displayed associations with carriers of *BRCA1* and *BRCA2* (Kuchenbaecker et al., 2015; Ramus et al., 2012; Ramus et al., 2011). One locus was identified that associated with *BRCA1* mutation carriers only (Couch et al., 2013). This suggests that low risk common variants interact multiplicatively with high risk rare variants in susceptibility to EOC (Kuchenbaecker et al., 2015). This then led to a GWAS meta-analyses of data from ovarian cancer cases unselected for family history, ovarian cancer cases with *BRCA1* mutations, and ovarian cancer cases with *BRCA2* mutations, which retrieved additional EOC susceptibility loci (Kuchenbaecker et al., 2015).

The EOC susceptibility loci identified so far explain 3.9% of the excess familial risk of EOC in the general population, 5.2% in *BRCA1* carriers, and 9.3% in *BRCA2* mutation carriers (Kuchenbaecker et al., 2015). The identification of these loci may be useful for risk assessment in individuals who carry mutations in *BRCA1* and *BRCA2,* yet their contribution to risk in the general population is still too low to be useful. Yet, these studies provide a starting point for the discovery of pathways and mechanisms operating in ovarian oncogenesis. Delineation of these pathways may reveal novel therapeutic strategies, much in the same way the identification of *BRCA1* and *BRCA2* and their role in homology-directed recombination led to the use of synthetic lethality with PARP1 inhibition in breast and ovarian cancer (Fong et al., 2009). However, the mechanistic underpinnings of these susceptibility loci remain largely unknown.

Functional Analysis of Susceptibility Loci

Several studies have performed functional analysis of susceptibility loci with different degrees of depth of investigation on how these susceptibility loci influence disease predisposition. From the experience of previous functional analysis of GWAS loci, described here is a thorough and potential flow of questions to analyze a locus (Figure 5). To begin functional analyses of the mechanisms of susceptibility of disease loci, all correlated SNPs must be identified since tagging SNPs are not necessarily the functional SNP, rather they are a surrogate marker for the locus (Carlson et al., 2004; Gabriel et al., 2002). Correlated SNPs can be retrieved using an arbitrary threshold of LD by obtaining data from the 1000 Genomes Project and performing haplotype analysis with the Haploview program (Barrett et al., 2005). Interestingly the majority of associated and correlated SNPs reside in non-coding regions of the genome (Maurano et al., 2012). Since these SNPs do not disrupt the amino acid sequence of proteins in the cell to disrupt cellular processes, it is hypothesized that associated SNPs exert their effects through changing the transcription activity of enhancers and promoters and therefore affecting the transcription rates of target genes (Freedman et al., 2011).

Fortunately, The Encyclopedia of DNA Elements (ENCODE) discovered that 80% of the genome has a biological function in at least one cell type (Dunham et al., 2012). Many of these functions are cell type specific indicating the importance of choosing the cell type for functional analyses (Heintzman et al., 2009).

After identifying which SNPs localize within a regulatory element, the next step is to identify which SNPs have allele specific activity. This is then followed by identifying

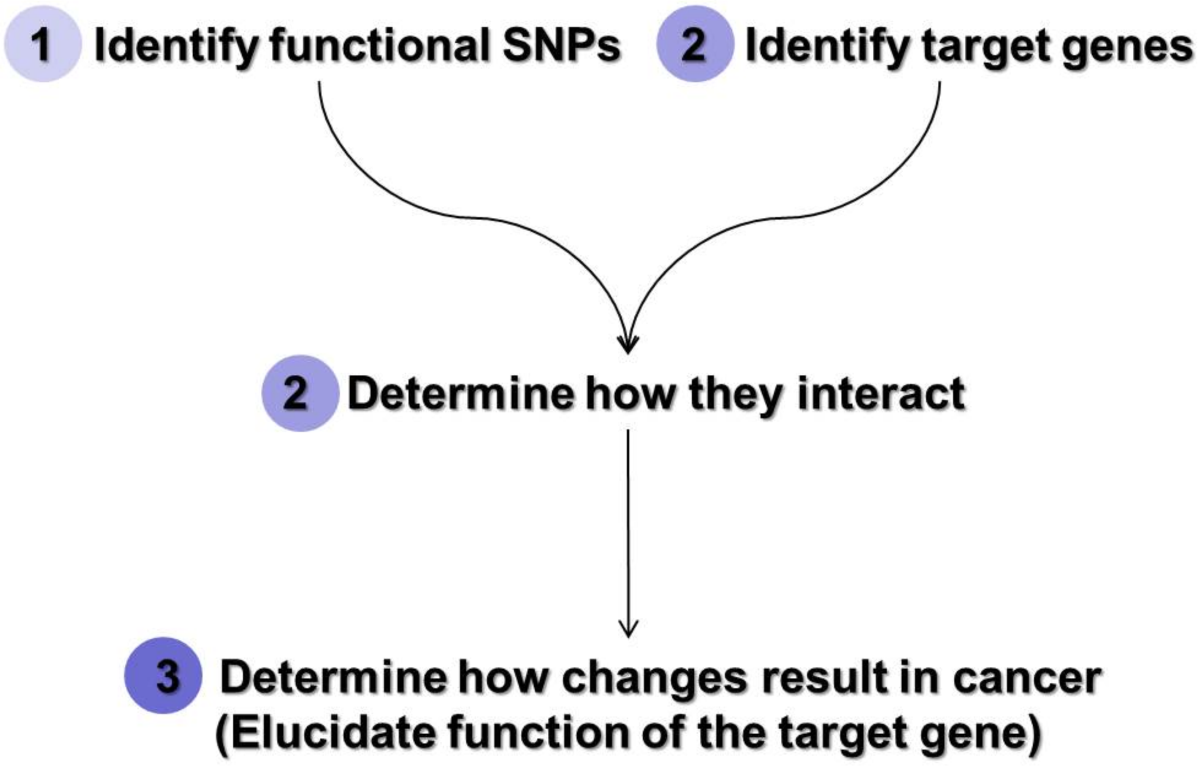

1 Identify functional SNPs
2 Identify target genes
2 Determine how they interact
3 Determine how changes result in cancer
(Elucidate function of the target gene)

Figure 5. Flow chart for the functional analysis of cancer susceptibility loci.

the downstream target gene whose transcription is affected by allele changes at the causal SNP (Figure 5). Genes within an arbitrary distance of the causal SNP can constitute the universe of candidate target genes for functional analysis. Within 1 megabase (Mb) of the causal SNP is within reason since very few enhancers loop to genes at a farther distance although, it is possible (Jin et al., 2013).

The next step would be to examine ways to link SNPs and genes (Figure 5). The final step would be to explore how these changes in alleles and expression of target genes result in disease, or in this discussion cancer. This last and final step would be an ongoing exploration and depending on the gene and pathways disrupted, would require different scientific techniques to analyze (Figure 5). Since risk variants identified by

GWAS, individually, contribute a small percentage of risk, presumably, these variants will have small effects on biological functions. Ultimately, functional analyses of GWAS loci uncover the mechanisms by which genetic variation influences risk.

Transcriptional Regulation

Since it is hypothesized that allele changes of GWAS SNPs most likely effect the transcription function of enhancers and promoters, a clear understanding of and development of functional assays is needed to identify the causal SNP. Regulation of transcription determines cell identity in development and maintains homeostasis of the cell. Aberrant transcription can lead to severe changes in the biological processes of a particular cell. Transcription is quite complex in that it is under combinatorial control (Britten and Davidson, 1969). Different genes are regulated by different combinations of TFs in a cell type specific manner (Britten and Davidson, 1969). Therefore, functional analysis of GWAS hits requires identification of a transcription regulatory network at a locus and where the glitch in the network, caused by the associated SNP, resides.

Basics of Transcriptional Regulation

There are three main DNA regulatory elements in transcription. The core promoter resides immediately upstream and adjacent to the transcription start site (TSS). General TFs bind to the core promoter. The general TFs make up a group of proteins that have or support the catalytic processes necessary for transcription elongation. The next regulatory element is the regulatory promoter and it lies immediately upstream of the core promoter. It recruits co-activating/repressing TFs that help recruit and activate or block and repress general TFs to the core promoter. The final regulatory element is the enhancer which lays several kilobases (kb) upstream or downstream of the core promoter. It loops to the core and regulatory promoter to aid in recruiting co-activating/repressing TFs. Enhancers also seem to be cell type specific and make up the majority of regulatory elements in the genome (Thurman et al., 2012). The core promoter is generally inactive in vivo without the regulatory promoter and/or enhancer. Since enhancers make up the majority of regulatory elements in the genome and have a greater influence on transcription activity, causal SNPs have a higher probability of influencing the function of an enhancer.

For DNA to fit and function in the nucleus, it has a specific structure and conformation specific to particular cell types. DNA is packaged into nucleosomes by wrapping around histone octamers which in turn package into chromatin. These histones also regulate which regions of DNA are accessible to TFs since chromatin needs to be de-condensed for activation at promoters and enhancers by specific TFs (Cairns, 2009). Chromatin loops bring together enhancers and promoters and organize the chromatin into areas of euchromatin (active) and heterochromatin (in-active).

The transcription cycle can be described in several steps (Reviewed in (Fuda et al., 2009). The first step involves clearing of nucleosomes from enhancers and promoters by specific TFs with the ability to bind to nucleosome bound DNA or at linker DNA between nucleosomes (Hebbar and Archer, 2003). Additional nucleosome remodeling TFs are recruited to further make DNA accessible. The second step involves the binding of co-activators which in turn recruit general TFs and RNA polymerase II to the core promoter. In the third step, DNA begins to unwind and RNA polymerase II initiates transcription. In the fourth step, co-activators phosphorylate RNA polymerase II which then escapes the core promoter and pauses. The fifth step requires further phosphorylation of RNA polymerase II to elongate and continue transcription of the gene. In the sixth step, RNA polymerase II elongates through the whole entire gene. In step seven, transcription terminates. In step eight, transcription reinitiates. All of these steps require TFs outside of the general TF category for continuation of the transcription cycle.

Once the appropriate TFs for the particular gene have activated enhancers and promoters, transcription elongation can begin. It has recently come to light that divergent RNA transcripts exist at promoters and enhancers (Core et al., 2014). Activation at an enhancer promoter interaction recruits the general TF machinery to anti-sense DNA as well as both strands of the enhancer creating unstable RNA (Core et al., 2014). This further portrays that specificity does not come from the core promoter where general TFs bind; rather it comes from the binding of regulatory TFs that activate transcription. Also, events downstream of transcription play an important role in gene expression. It seems that splicing determines whether RNA becomes stable or unstable. Thus, differences in enhancer RNA/anti-sense RNA and stable RNA is that stable transcripts have a binding motif for the U1 splicing complex which allows them to go through the translation process while unstable transcripts have a binding motif for the polyadenylation-dependent termination machinery (Core et al., 2014).

Transcription Factor Binding

TF binding can activate transcription by recruiting the general TFs to the core promoter or recruit chromatin remodeling enzymes that de-condense the chromatin to provide TFs access to the DNA (Blackwood and Kadonaga, 1998; Lee and Young, 2000; Struhl, 1998; Workman and Kingston, 1998). Alternatively, protein binding to the DNA can also lead to repression of transcription by competing for an activator's binding site, interacting with an activator or general TF and inhibiting their function, or by recruiting chromatin remodeling enzymes that condense the chromatin at a promoter or enhancer (Hanna-Rose and Hansen, 1996; Lee and Young, 2000; Struhl, 1998; Workman and Kingston, 1998). Different and distinct combinations of TFs regulate specific genes, repress or activate, and bind proximal and distal regulatory elements (Gerstein et al., 2012).

TFs typically have DNA binding domains that recognize specific DNA sequences. There are more than eighty known types of sequence specific DNA binding domains and those domains with similar amino acid sequences will bind to similar DNA sequences (Weirauch and Hughes, 2011; Weirauch et al., 2014). These TF binding sites are conserved in the genome and TF binding exhibits allele specific activity (Gerstein et al., 2012; Neph et al., 2012).

In vitro binding assays can identify the consensus sequence of many TFs. Yet in order for TFs to bind to their consensus sequence in the cell, the region may need to be accessible or nucleosome free, or their may need to be co-binding with other TFs. Additionally the TF needs to be expressed in the cell type. TF expression correlates to activity of enhancers and promoters with a TF motif (Ernst et al., 2011). ChIP-seq for

TFs identifies binding sites in the cell, but does not have the resolution of the in vitro assays to identify the exact binding sequence (Figure 6). A combination of both assays can inform the DNA binding properties of TFs. EMSA test allele specific activity of TF binding. Nuclear extracts mixed with radiolabeled probes containing the major or minor allele are run on a polyacrylamide gel (Kerr, 1995). If the allele disrupts a TF binding motif, the gel will show different band patterns between the major and minor allele probes indicative of changes in TF binding.

Chromatin Structure

As mentioned earlier in the text, histones play a major role in transcription regulation since they determine the accessibility of the DNA. Chromatin modifiers affect histone binding therefore specific histone modifications inform the activity of the DNA bound to that histone (Narlikar et al., 2002). Histone H3 lysine 4 tri-methylation (H3K4me3) marks DNA at promoters (Bernstein et al., 2005; Dunham et al., 2012; Ernst et al., 2011; Guenther et al., 2007; Heintzman et al., 2007; Mikkelsen et al., 2007). Increasing levels of transcriptionally engaged RNA polymerase II, or transcripts that consistently and stably transcribe, have increased levels of H3K4me3 (Core et al., 2014). Histone H3 lysine 4 di-methylation (H3K4me2) marks promoters and enhancers (Bernstein et al., 2005; Ernst et al., 2011; Heintzman et al., 2007). Histone H3 lysine 4 tri-methylation (H3K4me1) marks DNA at enhancers (Dunham et al., 2012; Ernst et al., 2011; Heintzman et al., 2007). Histone H3 lysine 27 acetylation (H3K27ac) and Histone H3 lysine 9 acetylation (H3K9ac) marks transcriptionally active DNA (Dunham et al., 2012; Ernst et al., 2011; Heintzman et al., 2007). Histone H3 lysine 36 tri-methylation and Histone H4 lysine 20 mono-methylation mark transcribed DNA (Ernst et al., 2011; Guenther et al., 2007; Mikkelsen et al., 2007). Histone H3 lysine 27 tri-methylation (H3K27me3) marks repressed DNA (Dunham et al., 2012; Ernst et al., 2011; Heintzman et al., 2007; Mikkelsen et al., 2007). Combinations of these histone marks inform whether the regions are active, weak or poised (Ernst et al., 2011). Promoter and enhancer states vary between active, weak and poised among different cell types but the regions of the DNA seem to maintain regulatory potential (Ernst et al., 2011). Promoters active in one cell type and not another seem to be cell type specific genes while promoters active in many cell types are metabolic housekeeping genes (Ernst et al., 2011). Enhancer locations are much more cell type specific than promoters. A gene active in more than one cell type uses one promoter yet uses a different enhancer in each cell type (Ernst et al., 2011). ChIP-seq for histone posttranslational modifications identifies the above mentioned regulatory regions (Figure 6). Due to their cell type specificity, to identify enhancers that affect a particular disease, the appropriate cell line must be used.

The opposite of histone binding is regions of open chromatin. Regions of open chromatin represent regions accessible to TFs and represent enhancers and promoters as well as insulators, silencers, and locus control regions (Gaszner and Felsenfeld, 2006; Li et al., 1999; Thurman et al., 2012). Thirty percent of distal open chromatin regions have marks of enhancers therefore the remaining sequences may contribute to chromatin organization (Heintzman et al., 2007). FAIRE identifies regions of open chromatin and therefore identifies regions with a potential biological function (Figure 6). FAIRE utilizes phenol chloroform extraction to separate the DNA, that tightly crosslinks to nucleosomes, from the unbound DNA in the aqueous layer (Giresi et al., 2007). Similar to TF binding, chromatin state differentially binds between maternal and paternal alleles (Dunham et al., 2012; Ernst et al., 2011).

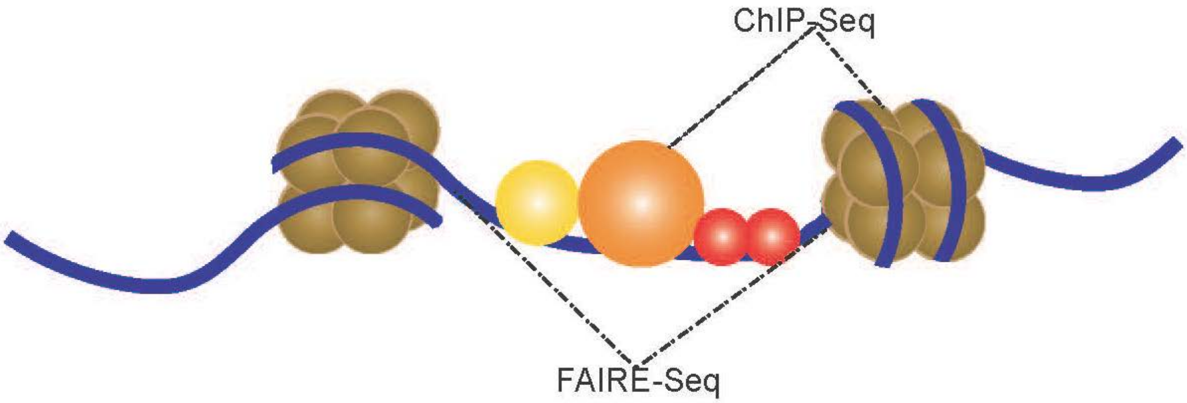


Figure 6: Regulatory regions and how they are identified. Regulatory regions are regions of open chromatin, bound by TFs. Modified histones also mark the transcription activity of DNA. FAIRE-seq identifies regions of open chromatin and ChIP-seq identifies TFs and modified histones bound to DNA.

3D Structure of the Genome

The genome is not structured in a straight line, rather, it is made of chromatin loops bringing together enhancers and promoters in transcription factories (Gondor and Ohlsson, 2009). 60% of promoters associate with one enhancer while 90% of enhancers associate with one promoter (Zhang et al., 2013). Additionally these interactions seem to be cell type specific with 60% of interactions occurring in only one cell line (Sanyal et al., 2012). It seems that each cell type utilizes a different enhancer for the same promoter (Ernst et al., 2011). Enhancers can loop to target genes up to 1

Mb away, in rare cases even farther (Jin et al., 2013). The average distance between interacting promoters and enhancers is 120 kb with almost half of the enhancer interactions occurring with the nearest promoter (Sanyal et al., 2012).

These enhancer and promoter interactions seem to be pre-formed in development and subsequent events occur to activate transcription rather than having dynamic interaction between enhancers and promoters to activate transcription (Ghavi-Helm et al., 2014; Kulaeva et al., 2012). Interestingly, there seems to be specific pairing between TFs at enhancers and promoters (Thurman et al., 2012). Possibly specific TFs for a developing cell work together to pre-form these interactions. Promoter and enhancer interactions occur at open chromatin sites and not at repressed sites (Sanyal et al., 2012; Thurman et al., 2012). Looped enhancers and promoters are bound by paused polymerase and transcription initiates once polymerase is activated for elongation through recruitment of additional factors (Ghavi-Helm et al., 2014). Promoters that do not interact with an enhancer have low gene expression levels (Zhang et al., 2013). Most likely, genes are deemed active during development.

Chromatin conformation capture (3C) identifies enhancer promoter interactions (Dekker, 2006). In this technique, formaldehyde cross links genomic DNA in live cells to maintain DNA interactions after cell lysis. Digestion and ligation of genomic DNA produces linear DNA products of the enhancer and promoter that can be amplified with specifically designed primers for those regions (Dekker, 2006).

Important, yet less studied, regulatory regions in maintaining the 3D structure of the genome are insulators and nuclear matrix/scaffold attachment regions (S/MAR). Both often mark borders between condensed and de-condensed chromatin

(Gerasimova et al., 2000). Insulators are regions of open chromatin bound by CTCF, and Cohesin creating chromatin loops that either block a promoter and enhancer interaction or maintain a promoter and enhancer interaction (Bondarenko et al., 2003; Kagey et al., 2010; Kulaeva et al., 2012; West et al., 2002). S/MARs are often a region of open chromatin that binds to the nuclear scaffold and surrounded by CTCF and H3K27me3 binding (Dunham et al., 2012; Guelen et al., 2008; Keaton et al., 2011; Mirkovitch et al., 1984; Ohlsson et al., 2001). These regions are also thought to be important in preventing aberrant enhancer activity and maintaining the 3D structure of chromosomes (Bushey et al., 2008; Guelen et al., 2008; Keaton et al., 2011; Linnemann et al., 2009). Reporter assays can identify whether an open chromatin region functions as an insulator or S/MAR by cloning the region in between an enhancer and promoter and observing decreased or lack of transcription (Kellum and Schedl, 1992). Isolation of the nuclear scaffold and DNA digestion and extraction can also identify S/MARs (Dijkwel and Hamlin, 1999; Keaton et al., 2011; Mirkovitch et al., 1984)

CHAPTER TWO:

IDENTIFYING FUNCTIONAL SNPS

Note to Reader

Two manuscripts that have been submitted for review include portions of this chapter.

Introduction

To establish the mechanism by which changes in alleles of SNPs in non-coding DNA contribute to EOC, we conducted a functional dissection of the 9p22.2 ovarian cancer susceptibility locus. The most significant SNP for high grade serous EOC reported initially (rs3814113; $P = 2.5 \times 10^{-17}$) is located 44 kb upstream and 220 kb downstream of the *BNC2* and *CNTLN* TSS, respectively (Song et al., 2009). The minor allele [C; MAF = 0.323] was associated with reduced risk of EOC (combined data OR = 0.82; 95%CI = 0.79-0.86). A total of twelve genotyped SNPs within the same linkage disequilibrium (LD) region ($r^2 \geq 0.239$; $D' \geq 0.591$) reached genome wide significance ($P < 5 \times 10^{-8}$) and mapped to non-coding regions, eight of which are located within intron 2 of the *BNC2* gene (Song et al., 2009).

Here, we conducted a comprehensive functional analysis of all SNPs at the locus in LD with rs3814113 ($r^2 > 0.3$). Since all of these SNPs fall in non-coding regions we hypothesized that they modify the activity of transcription regulatory elements present in enhancers and promoter regions (Freedman et al., 2011; Monteiro and Freedman, 2013) as these regions are the most common regulatory elements and thoroughly characterized. We integrated several methods to identify functional SNPs with allele-specific effects on enhancers and promoters operating in OSE and fallopian tube epithelial (FTE) cells (Coetzee et al., 2015).

Results

Candidate Functional SNPs

In order to identify a comprehensive set of candidate functional SNPs at the 9p22.2 locus, we downloaded all SNPs within 250 kb in LD ($r^2 \geq 0.3$) with rs3814113 in the 1000 Genomes Database (http://www.1000genomes.org/) using HaploView (Barrett et al., 2005). These 134 SNPs are distributed over an 82 kb region ranging from the first intron of *BNC2* to ~44 kb upstream from its TSS (Figure 8A and Table 1).

Table 1: Candidate Functional SNPs

Chromosome Position	SNP	r^2
chr9:16914834-16914835	rs10810671	1
chr9:16915020-16915021	rs3814113	1
chr9:16914894-16914895	rs7032221	1
chr9:16910676-16910677	rs10738467	0.974
chr9:16910762-16910763	rs10738468	0.974
chr9:16912987-16912988	rs4246134	0.974
chr9:16911637-16911638	rs4366169	0.974
chr9:16911756-16911757	rs4445329	0.974
chr9:16913042-16913043	rs4465052	0.974
chr9:16913285-16913286	rs4631563	0.974
chr9:16913472-16913473	rs6475092	0.974
chr9:16913513-16913514	rs6475093	0.974
chr9:16913615-16913616	rs6475094	0.974
chr9:16910897-16910898	rs7045767	0.974
chr9:16909050-16909051	rs7866677	0.949

Table 1 (Continued)

Chromosome Position	SNP	r^2
chr9:16914702-16914703	rs7048397	0.948
chr9:16916692-16916693	rs10962693	0.922
chr9:16908168-16908169	rs55689948	0.922
chr9:16907583-16907584	rs113780397	0.818
chr9:16911664-16911665	rs10465044	0.719
chr9:16857402-16857403	rs10962643	0.719
chr9:16905440-16905441	rs10962679	0.681
chr9:16909109-16909110	rs7851204	0.672
chr9:16891646-16891647	rs10124837	0.638
chr9:16889936-16889937	rs10962662	0.638
chr9:16911399-16911400	rs10810665	0.626
chr9:16911665-16911666	rs10810666	0.626
chr9:16912434-16912435	rs10810668	0.626
chr9:16912660-16912661	rs10810669	0.626
chr9:16912662-16912663	rs10810670	0.626
chr9:16915104-16915105	rs10962691	0.626
chr9:16913556-16913557	rs12377389	0.626
chr9:16913767-16913768	rs12377421	0.626
chr9:16910213-16910214	rs62543587	0.626
chr9:16913835-16913836	rs74664507	0.622
chr9:16858083-16858084	rs10756819	0.611
chr9:16878615-16878616	rs10756823	0.61
chr9:16878492-16878493	rs10122763	0.607
chr9:16876282-16876283	rs10810652	0.607
chr9:16877137-16877138	rs10810655	0.607
chr9:16894139-16894140	rs10962668	0.607
chr9:16881876-16881877	rs12345776	0.607
chr9:16887365-16887366	rs3927680	0.607
chr9:16882915-16882916	rs4644350	0.607
chr9:16881372-16881373	rs7040151	0.607
chr9:16915873-16915874	rs10962692	0.606
chr9:16909001-16909002	rs12376998	0.596
chr9:16873550-16873551	rs10810650	0.589
chr9:16874611-16874612	rs10962650	0.589
chr9:16900694-16900695	rs28498684	0.589
chr9:16900764-16900765	rs36116821	0.589
chr9:16863363-16863364	rs62541919	0.589
chr9:16891560-16891561	rs10962664	0.586
chr9:16891589-16891590	rs10962665	0.586
chr9:16892271-16892272	rs10962666	0.586
chr9:16914715-16914716	rs62543619	0.586
chr9:16856882-16856883	rs1416742	0.572
chr9:16914577-16914578	rs62543618	0.566
chr9:16898118-16898119	rs10962672	0.563
chr9:16896587-16896588	rs10962670	0.555
chr9:16848789-16848790	rs1339552	0.555
chr9:16903361-16903362	rs34606230	0.555

Table 1 (Continued)

Chromosome Position	SNP	r^2
chr9:16899284-16899285	rs62543561	0.555
chr9:16904634-16904635	rs62543578	0.555
chr9:16905169-16905170	rs62543579	0.555
chr9:16906005-16906006	rs62543581	0.555
chr9:16906093-16906094	rs62543582	0.555
chr9:16906151-16906152	rs62543583	0.555
chr9:16906306-16906307	rs62543584	0.555
chr9:16906888-16906889	rs62543585	0.555
chr9:16906509-16906510	rs72713890	0.555
chr9:16853778-16853779	rs10810647	0.541
chr9:16851677-16851678	rs4961501	0.541
chr9:16847519-16847520	rs7046326	0.541
chr9:16851976-16851977	rs7868157	0.541
chr9:16864520-16864521	rs2153271	0.539
chr9:16884585-16884586	rs10810657	0.528
chr9:16907645-16907646	rs181552334	0.527
chr9:16908382-16908383	rs80039758	0.516
chr9:16885016-16885017	rs12350739	0.508
chr9:16901227-16901228	rs13300853	0.501
chr9:16881255-16881256	rs7025549	0.481
chr9:16901066-16901067	rs62543565	0.475
chr9:16903365-16903366	rs10738466	0.451
chr9:16865698-16865699	rs12379183	0.445
chr9:16907966-16907967	rs117224476	0.44
chr9:16862279-16862280	rs7861573	0.426
chr9:16904947-16904948	rs10756835	0.425
chr9:16905327-16905328	rs12344726	0.425
chr9:16903947-16903948	rs7029285	0.425
chr9:16904079-16904080	rs7032175	0.425
chr9:16904201-16904202	rs7032420	0.425
chr9:16904354-16904355	rs7032581	0.425
chr9:16904140-16904141	rs7032644	0.425
chr9:16904302-16904303	rs58691828	0.425
chr9:16904495-16904496	rs7033061	0.425
chr9:16904640-16904641	rs7033084	0.425
chr9:16904704-16904705	rs7033194	0.425
chr9:16904845-16904846	rs7033354	0.425
chr9:16906358-16906359	rs7868583	0.425
chr9:16907996-16907997	rs77795022	0.422
chr9:16846259-16846260	rs1339547	0.391
chr9:16846322-16846323	rs1339548	0.391
chr9:16907674-16907675	rs76718132	0.379
chr9:16903849-16903850	rs10810661	0.37
chr9:16849603-16849604	rs10962641	0.37
chr9:16845725-16845726	rs10810645	0.346
chr9:16843012-16843013	rs4961498	0.336
chr9:16870181-16870182	rs62541922	0.317

Table 1 (Continued)

Chromosome Position	SNP	r^2
chr9:16895577-16895578	rs12551733	0.301
chr9:16902523-16902524	rs4961503	0.301
chr9:16907620-16907621	rs9697099	0.301
chr9:16846110-16846111	rs10962638	0.3
chr9:16863542-16863543	rs10962645	0.3
chr9:16868379-16868380	rs10962647	0.3
chr9:16868957-16868958	rs10962648	0.3
chr9:16873534-16873535	rs10962649	0.3
chr9:16874877-16874878	rs10962652	0.3
chr9:16876735-16876736	rs10962653	0.3
chr9:16877787-16877788	rs10962656	0.3
chr9:16881345-16881346	rs10962658	0.3
chr9:16883317-16883318	rs10962659	0.3
chr9:16858568-16858569	rs11788047	0.3
chr9:16872322-16872323	rs11789875	0.3
chr9:16864075-16864076	rs11792249	0.3
chr9:16889022-16889023	rs12376099	0.3
chr9:16854366-16854367	rs12379687	0.3
chr9:16852452-16852453	rs62541877	0.3
chr9:16861204-16861205	rs62541878	0.3
chr9:16861507-16861508	rs62541879	0.3
chr9:16865290-16865291	rs62541920	0.3
chr9:16870500-16870501	rs62541923	0.3
chr9:16877422-16877423	rs62541926	0.3
chr9:16885463-16885464	rs77507622	0.3

Functional Analysis of SNPs

Now that a set of candidate functional SNPs have been identified through genetic means, molecular biology techniques identify SNPs with a biological function. Those SNPs with a biological function most likely affect disease. Here we map SNPs to regulatory elements marked by histone markers and open chromatin as well as measure their transcription activities in luciferase assays and EMSA.

Mapping SNPs to regulatory elements. Since all 134 SNPs in the candidate functional set are located in non-coding regions, multiple functional assays are needed to identify regions of transcriptional regulatory activity. First we integrated FAIRE-seq, and ChIP-seq for H3K27Ac and H3K4Me1, histone markers for active chromatin and enhancers, respectively from Coetzee et al. (Coetzee et al., 2015; Ernst et al., 2011; Heintzman et al., 2007). We generated chromatin landscape profiles at the 9p22.2 locus (Figure 8A) from cell lines postulated to be the origins of high grade serous ovarian cancer and serous ovarian cancer cell lines.

Table 2: Twenty-two SNPs associated with ovarian cancer risk overlap with areas of regulatory activity. SNPs highlighted in dark blue have an $r^2 > 0.8$ to rs3814113. SNPs in blue have an $r^2 = 0.5 - 0.8$. SNPs in light blue have an $r^2 = 0.3 - 0.5$.

Region	chr9 Coordinates Tile	SNP Name	Effect Allele	Reference Allele	R^2
1	16,837,392-16,838,723				
2	16,848,158-16,848,790				
3	16,850,432-16,851,014				
4	16,852,717-16,853,479				
5	16,857,377-16,857,907				
	T5	rs10962643	A	C	0.719
6	16,860,790-16,861,348				
	T6	rs62541878	T	A	0.3
7	16,863,768-16,874,127				
	T7.1	rs11792249	G	T	0.3
		rs2153271	T	C	0.539
	T7.2	rs62541920	A	G	0.3
		rs12379183	G	A	0.445
	T7.3	rs10962647	G	T	0.3
	T7.4 & T7.5	rs10962648	C	G	0.3
	T7.6	rs62541922	C	T	0.317
		rs62541923	A	C	0.3
	T7.7	rs11789875	A	G	0.3
	T7.8	rs10962649	T	C	0.3
		rs10810650	T	C	0.589
8	16,883,570-16,885,692				
	T8	rs10810657	A	T	0.528
		rs12350739	A	G	0.508
		rs77507622	G	A	0.3
9	16,899,790-16,900,338				
10	16,901,238-16,902,039				
11	16,907,559-16,908,180				
	T11	rs113780397	A	G	0.818
		rs9697099	A	T	0.301
		rs181552334	G	A	0.527
		rs76718132	T	C	0.379
		rs117224476	G	T	0.44
		rs77795022	G	T	0.442
12	16,915,387-16,915,739				

FAIRE and ChIP-seq profiles revealed twelve regions showing evidence for enhancer activity in at least one ovarian cell line (Figure 8A). Twenty-two candidate functional SNPs (Table 2) overlapped with five regions containing FAIRE or ChIP-seq features (Figure 8A).

Development of Enhancer Scanning method. Several SNPs overlapped with regions of activity at the 9p22 locus. To eliminate non-causal SNPs in an accelerated fashion, we developed a streamlined method that can systematically scan for regions with regulatory activity in a cell line of choice and can be scaled-up to study relatively large genomic regions (100-500 kb). The method takes advantage of online bioinformatics tools and combines a PCR-based generation of tiling clones spanning the region with high efficiency recombination cloning.

Experimental Design. The complete procedure is depicted in Figure 7. To reiterate, we start with a genomic region identified by a GWAS as associated with risk for a certain condition (Figure 7) and to reiterate we assume that the tagging SNP may not necessarily represent the functional SNP. Next, in order to capture the variation in the locus we can use linkage disequilibrium (LD) structure information and retrieve all SNPs in (LD) with the tagging SNP (Figure 7) (Carlson et al., 2004; Hazelett et al., 2014). As a rule of thumb, we retrieve sets of SNPs in LD in 1000 Genomes Project with $r^2 \geq 0.3$, ≥0.5, or ≥0.8. These thresholds are arbitrary and serve as a guideline for the investigator to decide which set to pursue further analysis appropriate for the resources and time available.

With the regions of interests marked by enhancers defined, we next generate tiles by PCR spanning the regions marked by enhancers with each tile of approximately

2 Kb (Figure 7). This size is a compromise between having to generate the least amount of tiles to cover a region and efficient and reproducible PCR amplification. Regions with repetitive sequences may need smaller tiles to facilitate amplification. The PCR primers used are designed to include *att*R sequences to mediate recombinational cloning into destination vectors, pGL3-LRF and pGL3LRR, that contain a luciferase gene driven by a basal promoter (Figure 7). Using the Gateway® Vector Conversion System, we converted the pGL3-Promoter vector (Promega) to a Gateway® destination vector by inserting a blunt-ended cassette of the *ccd*B gene and the chloramphenicol resistance gene flanked by *att*R1 and *att*R2 sites into a *Sma*I site of pGL3-Promoter vector. Two recombinant vectors, carrying the cassette in each orientation, pGL3-LRF and pGL3-LRR, are then used to clone individual tiling clones by recombination. Plasmids containing tiling clones in both orientations are then transfected in an appropriate cell type and activity of luciferase is measured and compared with the corresponding pGL3-LR destination vector (Figure 7).

Although enhancers are expected to operate independent of orientation and position relative to the target promoter (Khoury and Gruss, 1983), because the relative position of the binding site in relation to the promoter of the reporter may influence expression, tiles should be tested in both cloning orientations.

True functional SNPs are also expected to show allele specific differences in enhancer activity. Risk alleles may create or disrupt a specific binding site therefore it is recommended to test both alleles for each SNP. Different alleles can be introduced in the tiles using Quick Change PCR mutagenesis (Braman et al., 1996).

Choice of template, tile verification, and host cell. Once the region of interest is defined we generate tiles by PCR using human genomic DNA or a Bacterial Artificial Chromosome (BAC) clone containing the region of interest (Figure 7). In our experience the latter provides a more robust option to amplify the tiles. It is expected that the BAC is likely to contain the major allele for the SNPs, but could represent the minor allele. In addition, depending on the number of loci being studied and the size of regions to be analyzed there could be hundreds of tiling clones to be processed simultaneously

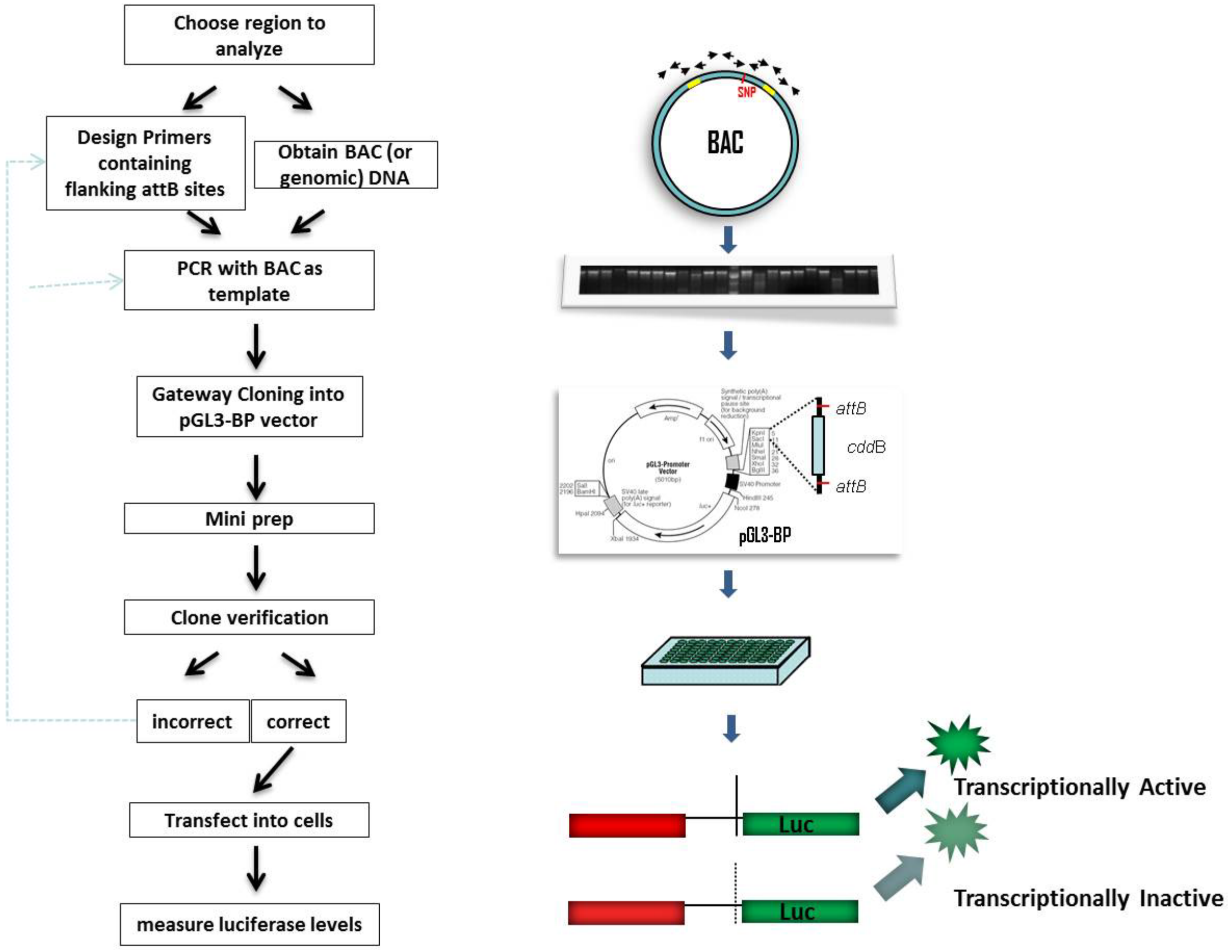

Choose region to analyze
Design Primers containing flanking attB sites
Obtain BAC (or genomic) DNA
PCR with BAC as template
Gateway Cloning into pGL3-BP vector
Mini prep
Clone verification
incorrect
correct
Transfect into cells
measure luciferase levels
SNP
BAC
Synthetic poly(A) signal / transcriptional pause site (for background reduction)
Ampr
f1 ori
ori
pGL3-Promoter Vector (5010bp)
KpnI 5
SacI 11
MluI 15
NheI 21
SmaI 28
XhoI 32
BglII 36
SV40 Promoter
HindIII 245
NcoI 278
luc+
SV40 late poly(A) signal (for luc+ reporter)
2202 SalI
2196 BamHI
HpaI 2094
XbaI 1934
attB
cddB
attB
pGL3-BP
Luc
Luc
Transcriptionally Active
Transcriptionally Inactive

Figure 7: Luciferase assay protocol. Here we designed a semi-high-throughput assay that utilizes transcription reporter plasmids to test several potential regulatory elements for transcription activity.

increasing the chance of sample mix-ups. Thus, it is useful to sequence a sample (or all) of the clones to confirm their identity and to determine the correct allele being tested.

Ideally the host cell for the transfection should represent a tissue compartment relevant for the disease under study. For example, cells for normal intestinal crypts when studying colorectal cancer. However, primary cells might be difficult to transfect and immortalized or cancer cells may provide an alternative, but the use of these cells lines should be kept in mind when interpreting the results.

Controls. Several controls are advisable to guarantee data quality and wise use of resources and reagents. We found that because in each experiment a large number of luciferase measurements are going to be made, it is important to have a monitor of whether the transfection has worked before processing samples. We recommend the use of a parallel transfection with a GFP expression vector of choice and when there is no detectable GFP-positive cell we do not proceed to lyse cells and measure luciferase. We also include a positive control (pGL3-Control Vector containing with a SV40 enhancer) and expect its activity to be consistent across transfections (10-20X the negative control in the experiment, for example). Eight replicates are performed for each tile in each orientation. The negative controls used are the plasmids that only include the recombination cassette in both orientations and are compared to the tested tiles of the same orientation. An additional control derived from the region being studied is also recommended and can be designed by identifying a region with no chromatin markers indicative of activity.

Statistical analysis. Results from transfections with individual tiles are analyzed in eight replicates. This number allows for the generation of box whisker plots and maximizes the 96-well setup but investigators might want to scale it down to reduce costs. Raw readings from firefly (*Photinus pyralis*) luciferase are normalized against an internal *Renilla reniformis* luciferase control driven by a minimal promoter, pRL-TK (*Renilla* luciferase driven by thymidine kinase promoter) to adjust for differences in transfection efficiency in different wells. Next, we set the negative control (pGL3-LRF and pGL3-LRR empty vectors) as the reference and results are transformed to fold over the negative control. Statistical analysis is performed by comparing the exact means and $p \leq 0.05$ is considered significant. When testing a large number of tiles, a multiple testing correction can be applied (for example, $p \leq 0.05$/number of tiles being tested). Alternatively, a less stringent false discovery rate can also be applied for prioritization. However, we feel that due to the stage of SNP analysis in which this assay is being performed it would be unwarranted to apply multiple testing corrections as true positive clones with small effects might be discarded. Additional more stringent tests are subsequently applied to the tiles (for example, allele-specific differences, activity in EMSA, etc.) to weed out false positive hits.

Anticipated results. At the end of the protocol we anticipate that the investigator will have generated tiles representing the genomic region of interest and will have identified those that contain regulatory regions capable of activating transcription of a heterologous reporter gene in the cell line of choice. These tiles can then be reduced to narrow down the region and they can be mutagenized to test whether different alleles of

SNPs contained in the tile have specific effects. The results provide an empirical map of regulatory activity operating in the locus for the cell lines tested.

Limitations. The enhancer scanning method presented here is based on the use of a plasmid-based reporter gene to detect regions in the genomic DNA with transcriptional regulatory activity. As other naked DNA-based assays (such as EMSAs) it identifies sequences present in the DNA that have the ability to recruit and bind transcriptional regulators. It is conceivable that the underlining DNA (exposed in the plasmid-based assay) that carries the activity may, in certain chromatin contexts, be hidden by tight nucleosome packing, repressive chromatin features and DNA methylation. Thus, some tiles may be false positive hits.

The sensitivity of the method presented here is dependent on transfection efficiency and cells that are difficult to transfect, or experiments in which transfection is sub optimal, may not identify all tiles with activity. Use of a set of internal controls greatly minimizes false negatives due to transfection failures. The ability of the enhancer to activate transcription of a cognate promoter region depends of the formation of an adequate DNA loop between the enhancer and promoter (Plank and Dean, 2014). Thus, it is conceivable that even when testing both cloning orientations in a plasmid context, an optimal loop may not form between the region and the promoter leading to false-negative results. Our data suggests that the fraction of false negatives due to the plasmid context is small.

Finally, it is unclear to what extent the detection of a regulatory activity using a heterologous promoter affects the results. It is possible that the enhancer-promoter interaction depends on binding factors or other promoter features (biochemical compatibility) (van Arensbergen et al., 2014) that are not found in the SV40 promoter. In this case, the enhancer scanning method may not properly identify the region. Thus, we suggest that strong candidates be tested also against the promoter of candidate target regions in the locus.

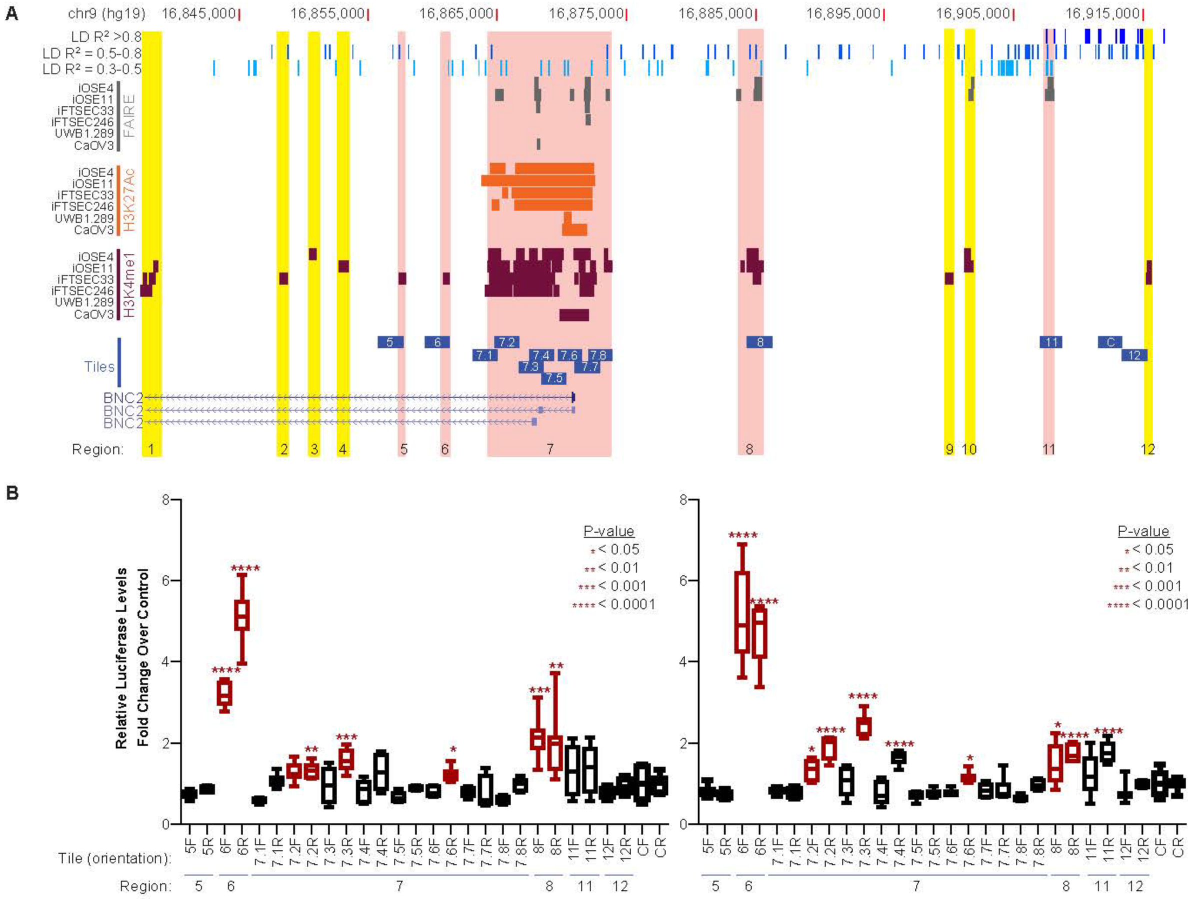


Figure 8: Functional annotation identifies candidate functional SNPs overlapping with regions of regulatory activity in ovarian cells. *A*. Within the region of the 9p22 locus containing linked SNPs, twelve regions contain FAIRE peaks (gray), H3K27Ac peaks (orange), and/or H3K4Me1 peaks (maroon) in iOSE, iFTSE, and ovarian cancer cells. Regulatory regions highlighted in yellow do not overlap with candidate functional SNPs. Regions highlighted in red overlap with candidate functional SNPs. Blue bars represent location of 2 kb tiles cloned into luciferase reporter vectors. ***B*.** Box and whisker plots show the luciferase activity from duplicate experiments with 8 biological replicates of each tile in both orientations. Asterisks denote tiles exhibiting significant transcription activity compared to a control tile (C) located in a genomic region in the locus inactive in ovarian cells as judged by features in the figure. Tiles moved forward in the functional assays are colored red.

Mapping SNPs to regions of enhancer activity. Using the Enhancer Scanning method, we tested twelve genomic tiles (~2 kb each) (Figure 8A), in both orientations, spanning the five functional regions cloned upstream of the SV40 promoter driving expression of luciferase. We also tested two additional tiles: one overlapping with the most significant SNP (T12.1) and a control tile devoid of enhancer activity as judged by FAIRE and ChIP-seq data (Figure 8A, Tile C).

Tiling clones were transfected in IOSE4^{cMYC} ovarian cells (an early stage in vitro transformation model of ovarian cancer) (Lawrenson et al., 2010) and luciferase levels were determined 24h post transfection. Tiles in regions 6 (T6), 7 (T7.2, T7.3, T7.6), and 8 (T8) containing 9 candidate functional SNPs displayed significant activity in either orientation (two tailed *t*-test p<0.05 compared to the control tile C, repeated in duplicate tests) (Figure 8B).

SNPs with Allele Specific Effects

SNPs with allele specific enhancer activity. We further hypothesized that SNPs likely to have a functional impact will display allele-specific effects. Thus, we performed site directed mutagenesis to switch from the reference to the effect allele in tiles with significant luciferase activity and compared the activity of different alleles. Significantly different activity between reference and effect allele was found for seven SNPs in three regions, T6 (rs62541878), T7 (rs62541920, rs12379183, rs1092647), and T8 (rs77507622, rs10810657, rs12350739), which were retained for further analysis (Figure 9A).

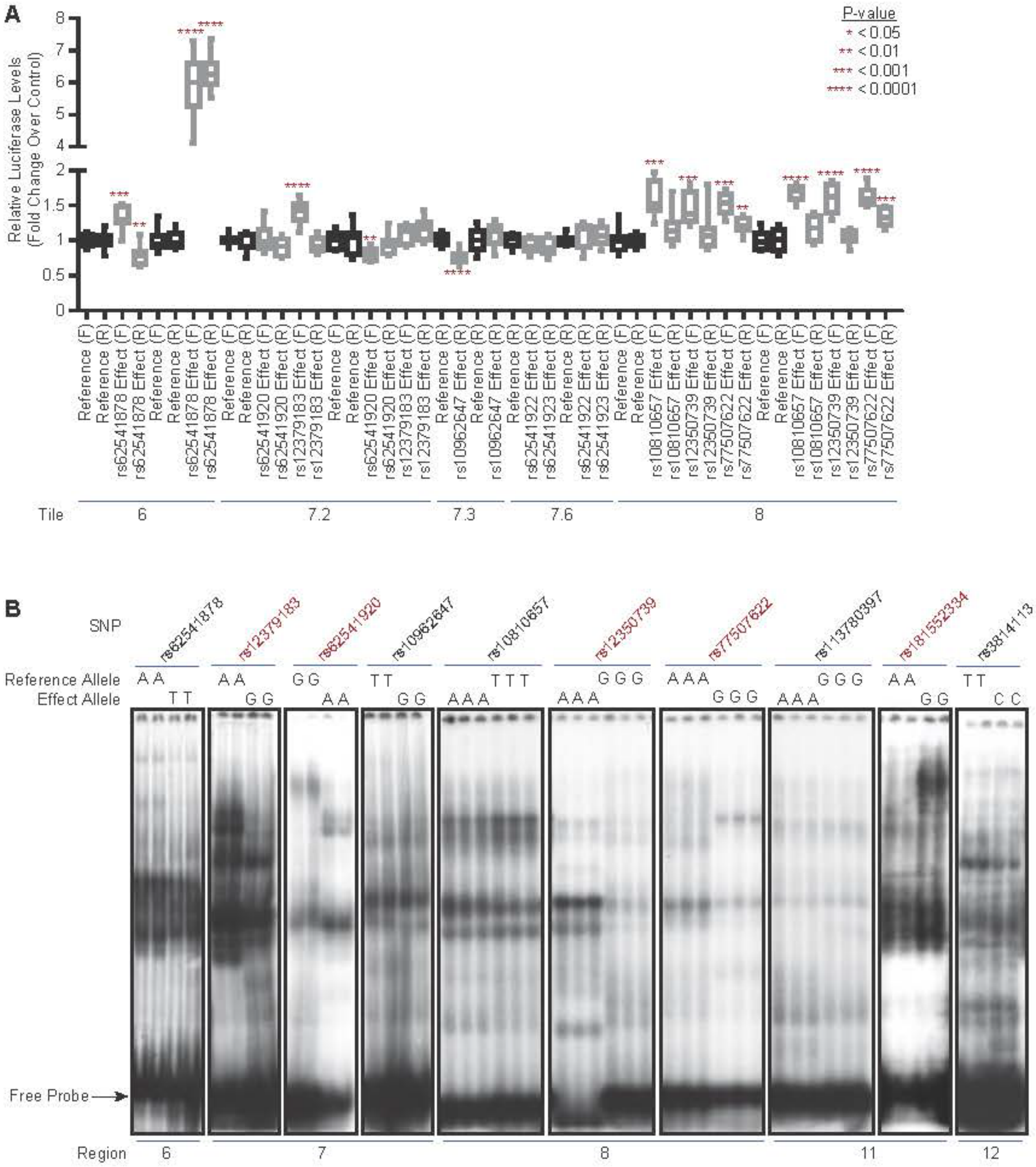


Figure 9: SNPs showing allele-specific activities. ***A.*** Luciferase assays reveal significant allele specific differences in transcription activation for rs62541878, rs62541920, rs12379183, rs1092647, rs77507622, rs10810657, and rs12350739 as indicated by red asterisks. Reference and effect allele tiles are shown as black and gray box and whiskers, respectively. ***B.*** EMSA showing allele specific differences in mobility between the reference and effect alleles. SNPs in Regions 7, 8, and 11 display differences in complex formation between the reference and effect alleles. SNPs with allele specific differences are indicated by red text.

Allele specific activities in electrophoretic mobility shift assays. We conducted electrophoretic mobility shift assays (EMSA) (Kerr, 1995) using fourteen 41-mer probes to interrogate both alleles for each of the seven SNPs listed above and located in regions 6, 7 and 8 (Figure 9B). Tile 11 had significant transcription activity in only one reporter experiment but two SNPs within the region (rs113780397 and rs181552334) are correlated with r^2 of 0.818 and 0.5 respectively, (the highest r^2 values of all candidate functional SNPs (Table 2)) and so four additional probes were tested. We also tested rs3814113, the most significant SNP, for allele-specific effects.
We obtained nuclear extracts from IOSE4^{cMYC} cells growing in log phase and incubated with oligonucleotide probes containing the reference or the effect allele. EMSA revealed allele specific effects for rs12379183, rs62541920 (Region 7), rs12350739, rs77507622 (Region 8) and rs181552334 (Region 11) (Figure 9B) indicating that five SNPs in three regions at the 9p22.2 locus are functionally relevant.

Summary

The two SNPs in Region 7 reside in an approximately 7kb region that includes the TSSs for two *BNC2* transcripts (Figure 11 and Figure 12A) denoted by FAIRE-seq and H3K4me1 ChIP-seq data in ovarian cells, and ENCODE layered H3K4me3 (promoters) ChIP-seq data (Figure 11 and Figure 12A). This region is the major *BNC2* promoter, implicating *BNC2* as a candidate mediator of ovarian cancer susceptibility at the 9p22.2 locus (Figure 11).

Region 8, containing two SNPs with allele specific activity in luciferase assays and EMSA, overlaps with FAIRE-seq and ChIP-seq data in ovarian cells with features

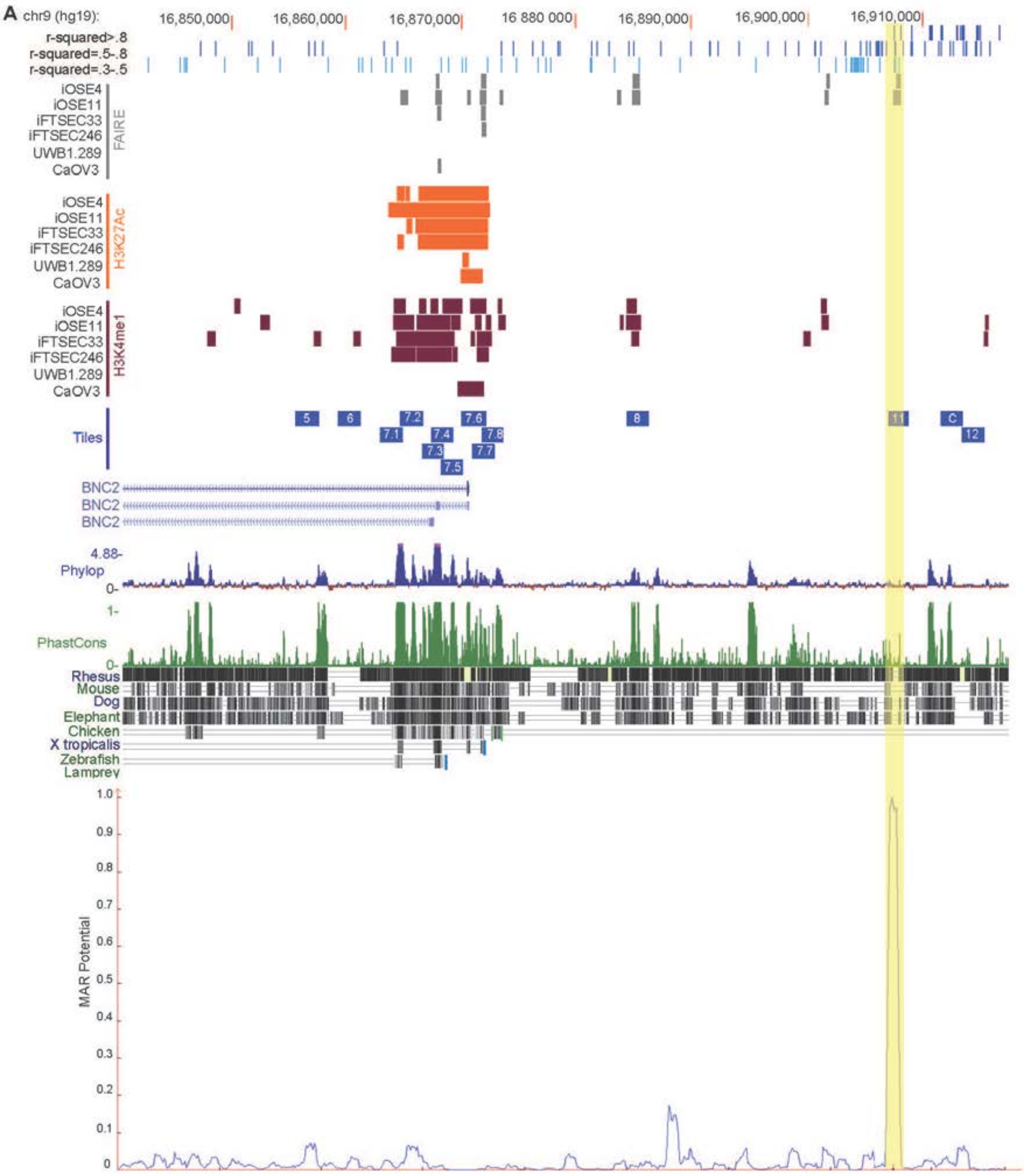


Figure 10: Conservation and S/MAR predicted sequences within the locus. A. This snapshot from the genome browser for the region containing linked SNPs includes tracks for Phylop, PhastCons scoring for conservation and alignment of DNA sequences among several vertebrates. Interestingly, region 7 and 8 have peaks of conservation for both scoring systems while regions 6 and 11 lack a conservation signal. **B.** Region 11 contains sequences highly predicted by MAR-Wiz to attach to the nuclear scaffold/matrix compared to the rest of the locus. The intensity of prediction was driven by the Origin of Replication Rule and the A-T Richness Rule.

indicative of an enhancer (Figure 11 and Figure 8A). Region 11 overlaps with FAIRE-seq data in ovarian cells and one SNP displayed allele-specific effects in EMSA experiments. Yet, it lacked ChIP-seq data for enhancer histone marks and had weak evidence for enhancer activity in luciferase assays. Interestingly, the region contains a sequence predicted to bind to the nuclear scaffold/matrix (Figure 11 and Figure 10) by MAR-Wiz (http://genomecluster.secs.oakland.edu/MarWiz/). The FAIRE-seq peak indicates lack of nucleosome formation at region 11 (Figure 10 and Figure 8A) consistent with observation of scaffold/matrix attachment regions (S/MARs) in plant and human cells (Keaton et al., 2011; Pascuzzi et al., 2014). The region is also A/T rich, a common feature of S/MARs (Keaton et al., 2011). Although S/MARs are poorly characterized functionally they have been suggested to regulate adjacent genes (Linnemann et al., 2009).

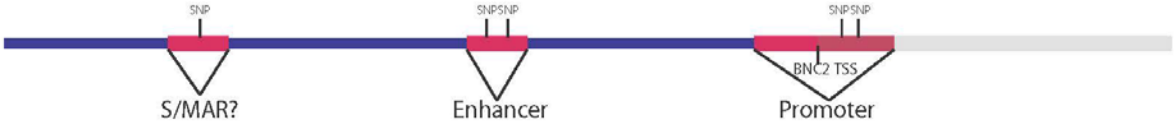


Figure 11: Summary of location of SNPs with most compelling evidence for function. Five out of 134 candidate functional SNPs have the most compelling evidence for function. Two reside within the promoter of *BNC2.* Two reside within an enhancer and one resides in a putative S/MAR.

We are continuing to investigate the potential S/MAR by performing a nuclear scaffold extraction (Dijkwel and Hamlin, 1999; Keaton et al., 2011) followed by qPCR for region 11. Additionally ChIP for Lamin B1, a nuclear matrix protein, and CTCF may also give additional evidence for S/MAR sites (Guelen et al., 2008; Keaton et al., 2011). Even though enhancers make up the majority of regulatory elements in the genome, other potential regulatory elements such as insulators and S/MARs need further characterization since they may affect disease such as we have shown in this chapter.

Materials and Methods

Candidate Functional SNPs

In order to identify a set of candidate functional SNPs in the locus we downloaded all SNPs within 250 kb of rs3814113, the SNP originally associated with ovarian cancer risk (Song et al., 2009) from the 1000 Genomes Project (v3) (Abecasis et al., 2012). The data was uploaded into Haploview. SNP retrieval was done by running Tagger only including rs3814113 and capturing all SNPs within 250 kb of rs3814113 resulting in 134 SNPs with an $r^2 > 0.3$.

Cell Lines and Cell Type-specific Datasets

The contribution of various cell and tissue types for the origin of different invasive EOC subtypes, and their molecular profiles indicate that histotypes of EOC should be considered different diseases (Berns and Bowtell, 2012). Subtype analysis of the association of the most significant SNP (rs3814113) revealed a stronger association when the analysis was restricted to high grade serous tumors, marginally associated with endometrioid tumors and no evidence of association was seen for mucinous or clear cell carcinoma, although non-serous subtypes have relatively small sample sizes. Thus, we chose to use, whenever possible, cell lines and datasets originating from ovarian surface and fallopian tube epithelium. Experiments were conducted in two immortalized normal ovarian surface epithelial cell lines, iOSE4 and iOSE11 (Lawrenson et al., 2009), two immortalized normal fallopian tube surface epithelial cells (iFTSEC33 and iFTSEC246). In addition, we used a normal epithelial ovarian cell line immortalized with *hTERT* and transformed with *MYC* called IOSE4^{cMYC} (Lawrenson et al., 2010) and two ovarian cancer cell lines, CaOV3 considered highly likely high grade serous carcinoma by molecular profiling (Domcke et al., 2013) and UWB1.289 (*BRCA1*-null)(Dellorusso et al., 2007). (Dellorusso et al., 2007).

Immortalized ovarian cells were grown in media with a 1:1 ratio of MCDB105/Medium 199, 15% Fetal Bovine Serum (FBS), 10 ng/mL Epidermal Growth Factor, 0.5 µg/mL hydrocortisone, 5 µg/mL insulin, and 34 µg/mL Bovine Pituitary Extract. For 4C2 cells medium was complemented with 2 µg/mL Blasticidin.

FAIRE-Seq and ChIP-Seq for Histone Modifications

FAIRE-Seq and ChIP-Seq for Histone H3 Lysine 27 Acetylation and Histone H3 Lysine 4 Monomethylation was performed in iOSE4, iOSE11, iFTSEC33, iFTSEC246, UWB1.289, and CaOV3 (Coetzee et al., 2015).

Enhancer Scanning

A series of genomic tiles of ~2 kb spanning regions with evidence of regulatory activity in experiments for FAIRE-Seq and ChIP-Seq for Histone modifications and containing significantly associated SNPs were generated by PCR amplification with KOD (Millipore) or Taq Polymerase (Qiagen) using 50 ng of bacterial artificial chromosome (BAC) Clone RPCI-11-185E1 (Empire Genomics) as template. Tiles were cloned in both a forward and reverse orientation upstream of the SV40 promoter by recombination in the firefly luciferase reporter vector pGL3-Pro-*attb* vector designed to test for enhancer regions. It is a modification of pGL3-Promoter (Invitrogen) adding *att*b sites surrounding the *cddb* gene.

Each tiling clone (100 ng) was co-transfected in eight replicates into IOSE4^{cMYC} cells with 10ng of pRL-CMV (Promega), an internal control expressing *Renilla* luciferase, per well of 96 well plates of IOSE4^{cMYC} cells. Luciferase activity was measured 24 hours post transfection by Dual Glo Luciferase Assay (Promega). Quick Change II XL Site Directed Mutagenesis Kit (Agilent) was used to mutate the tiles from the reference to the effect alleles.

Firefly Luciferase counts are normalized by *Renilla* luciferase counts in each sample. Each read is then divided by the average normalized read of TC (the control tile devoid of any activity-associated chromatin features) for scanning experiments, or by the normalized read of the plasmid with the reference allele in allele-specific experiments to generate the normalized fold change over the control. Tiles with significantly (two tailed t-test <0.05) higher luciferase counts than the control tile (TC) in two independent experiments in both orientations were tested for allele specific effects. Tile T7.2 was significant in only one experiment for the forward orientation but its reverse orientation was significant in both experiments and was thus included. For allele-specific luciferase assays, tiles with the effect allele were considered significant if the luciferase counts were significantly higher (p-value <0.05) in one independent experiment than the tile with the reference allele.

Genome Browser

Bed files using the hg19 genomic positions were created for the ChIP-seq and FAIRE-seq data (Coetzee et al., 2015), as well as a list of the candidate functional SNPs, and enhancer scanning tiles. These bed files were uploaded to the genome browser by clicking manage custom tracks followed by add custom tracks in our own session file of the genome browser. We could then visualize and overlap SNPs with the regulatory elements and design tiles that overlap with the regulatory elements containing SNPs. Under this session we can also include tracks from the genome browser.

Electrophoretic Mobility Shift Assays

Nuclear extracts were obtained from IOSE4^{cMYC} cells at 70-90% confluence. Cells were harvested, pelleted at 450 *g* for 5 minutes, and suspended gently in 1X Lysis Buffer (10 mM HEPES pH7.9, 1.5 mM $MgCl_2$, 10 mM KCl, 10 mM DTT, protease inhibitor cocktail), and incubated on ice for 15 minutes. The cell suspension was centrifuged for 5 minutes at 450 *g* and the pellet was re-suspended in 1X Lysis Buffer. The cells were disrupted using a syringe with a narrow gauge (No. 27) and nuclei were pelleted by centrifugation at 11,000 *g* for 20 minutes followed by re-suspension in 1X Extraction Buffer (20 mM HEPES pH7.9, 1.5 mM $MgCl_2$. 0.42 M NaCl, 75% Glycerol, 10 mM DTT, protease inhibitor cocktail) and shaken for 30 minutes. The nuclear extracts were cleared by centrifugation at 21000 *g*, 4°C.

Single stranded DNA (ssDNA) oligonucleotides containing the reference or effect allele in the center of the 41-mer probe were synthesized (Invitrogen). The probe is prepared by heating complementary ssDNA to 100°C in Elution Buffer (10 mM Tris·Cl, pH 8.5) and annealed by slowly decreasing the temperature. Annealed probes (10 pmol) were labeled with γ^{32}P-dATP using T4 polynucleotide kinase (NEB) for 1 hour at 37°C, followed by 30 minutes at 65°C. The unincorporated dNTPs were removed using the Qiaquick nucleotide removal kit (Qiagen). For each binding reaction 10 µg of nuclear extract were mixed with 1 µL of labeled probe in 1 X Binding Buffer (10 mM Tris, 50 mM KCl, 1 mM DTT; pH 7.5) with 1 µg poly dI-dC and incubated for 20 minutes, room temperature. Loading Buffer (5X) was added to the binding reaction and loaded in 6% native polyacrylamide gel. The gel was pre-run for at least 1 hour at 100V, loaded and electrophoresed at 80V overnight.

CHAPTER THREE:

IDENTIFICATION OF TARGET GENE

Note to Reader

A manuscript that has been submitted for review includes portions of this chapter.

Introduction

In this chapter, we next identified the major target gene regulated by the candidate functional SNPs at the 9p22.2 ovarian cancer susceptibility locus. We tested whether candidate target genes are functionally and physically associated with the previously identified functional SNPs. The closest gene to these SNPs is *BNC2*. *BNC2* expression has been compared between immortalized normal ovarian epithelial cell lines and ovarian cancer cell lines and *BNC2* expression decreases in cancer cell lines compared to normal (Goode et al., 2010). Additionally, immortalized ovarian epithelial cells subjected to transformation with c-MYC and KRAS also displayed reduced expression of *BNC2* compared to the parental cell lines (Goode et al., 2010) implicating *BNC2* as a potential target for ovarian cancer predisposition as well as a potential tumor suppressor. Since enhancers can loop to target genes at an average of 1 Mb away (Jin et al., 2013) only looking at the nearest gene would not be the most agnostic approach.

Results

BNC2 and CNTLN as Candidate Target Genes

Transcription activity at candidate gene promoters. Next, we examined all four genes (*c9orf92, BNC2, CNTLN* and *SH3GL2*) within 1 MB at either side of the region defined by the candidate SNPs (Figure 12A). We first identified their promoters by the presence of layered H3K4me3 in the vicinity of TSSs on seven non-ovarian cell lines from ENCODE (Integrated Regulation from ENCODE) (Figure 12A). The H3K4me3 promoter mark does not depend on cell type specificity as much as enhancers but rather promoters resides at similar locations across all cell types (Dunham et al., 2012). Next, we determined whether the gene was expressed in ovarian cell lines by examining the presence of H3K27ac as a surrogate marker of active promoters as well as transcript levels from RNA-sequencing data for ovarian and fallopian tube epithelial cells (Figure 12B). While marks in the *BNC2* and *CNTLN* promoters, combined with RNA-seq data indicate that they are expressed in ovarian cells, we saw little evidence of expression for *c9orf92* and *SH3GL2* (Figure 12B). Taken together, our results indicate that *BNC2* and *CNTLN* are the strongest candidate gene targets at this locus.

Expression Quantitative Trait Loci (eQTL). Measurement of *BNC2* and *CNTLN* mRNA levels in ovarian tissue binned by genotype (AA, AT, or TT) (eQTL) was performed to test whether the genotype at rs3814113 correlates to expression of genes within the 1 Mb region surrounding the SNP. There was no correlation between the SNP and expression changes of *BNC2* or the next closest gene, *CNTLN*. eQTL analysis was

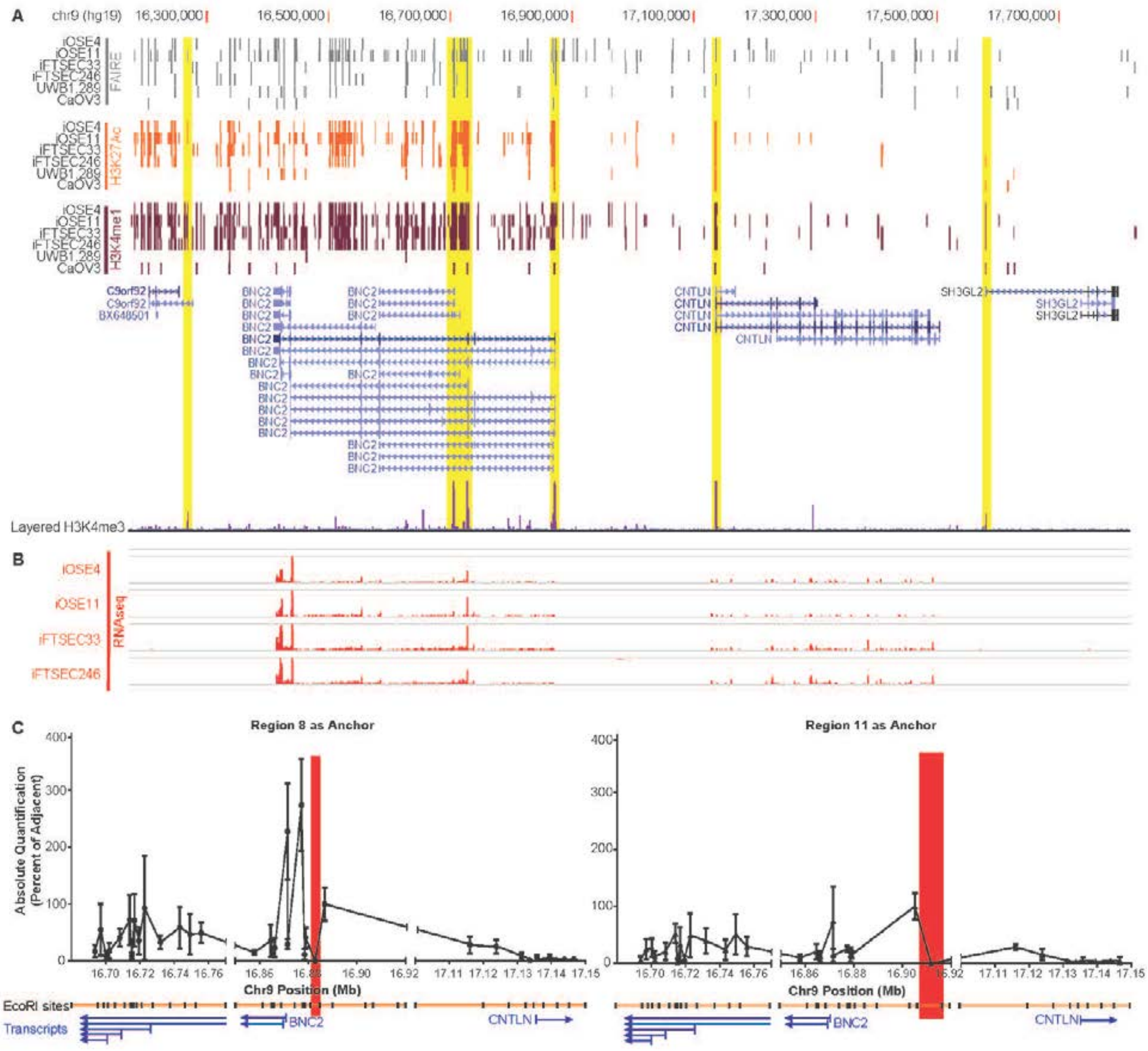


Figure 12. Functional SNPs influence transcription of *BNC2*. *A.* A snapshot from the genome browser displays UCSC genes as well as FAIRE peaks (gray), H3K27Ac peaks (orange), and/or H3K4Me1 peaks (maroon) in iOSE, iFTSE,, and ovarian cancer cells generated in the laboratory. The four genes within the region considered as potential target genes for ovarian cancer susceptibility include *c9orf92*, *BNC2*, *CNTLN*, and *SH3GL2*. ENCODE H3K4me3 peaks (purple), used to identify the promoters for these four genes (highlighted in yellow). H3K27ac tracks (orange) inform the extent to which these promoters are active and show that *BNC2* and *CNTLN* promoters are active in ovarian cells while *c9orf92* and *SH3GL2* are less active. ***B.*** RNA-seq for these four genes indicates the presence of transcripts for *BNC2* and *CNTLN* but not for *SH3GL2* and *c9orf92*. **C.** 3C analysis indicates that Region 8 (left) interacts with the *BNC2* promoter while region 11 (right) does not show a significant interaction compared to the adjacent site. Anchor regions for 3C are highlighted in red.

also performed using TCGA ovarian tumor data. In this dataset the genotype at rs3814113 does not correlate to expression of *BNC2* or *CNTLN*. eQTL was also performed genome wide among TCGA ovarian tumor data and not a single gene had expression level significantly correlating to the genotype. Since the SNP does not correlate to expression in normal or tumor tissue there may be context specific gene regulation of *BNC2* and cannot be visualized from these samples. A null eQTL has also been seen for TFs *MYC, ESR1, and KLF4* with breast cancer associated SNPs even though other methods point to these genes as likely targets (Li et al., 2013). It is presumed that TFs are tightly regulated. Therefore a positive eQTL will not be observed.

Region 8 is in Physical Proximity to the TSS of *BNC2* in Ovarian Cells

We used Chromatin Conformation Capture (3C) to identify which promoters in the locus interact with Region 8. In iOSE11 cells, Region 8 displays more frequent interactions with the canonical *BNC2* promoter compared to an adjacent restriction site (Figure 12C). Region 11 showed no significant interactions with the promoters for the canonical and alternative *BNC2* transcripts, or with the *CNTLN* promoter (Figure 12C). Taken together the data indicate that Region 7 and 8 are involved in the regulation of the transcription of *BNC2.* The modules in Region 7 affect the major promoter of *BNC2* and the module in Region 8 is a distal regulatory enhancer which physically interacts with the *BNC2* promoter.

Summary

In summary, we deduced from four candidate functional SNPs, that *BNC2* is the target gene of this locus. Transcription activity at promoters measured by ChIP-seq for H3K27ac and presence of transcripts measured by RNA-seq narrowed our candidates to *BNC2* and *CNTLN*. Finally, physical interactions between the enhancer and promoter of *BNC2* measured by 3C identified our final candidate (Figure 13).

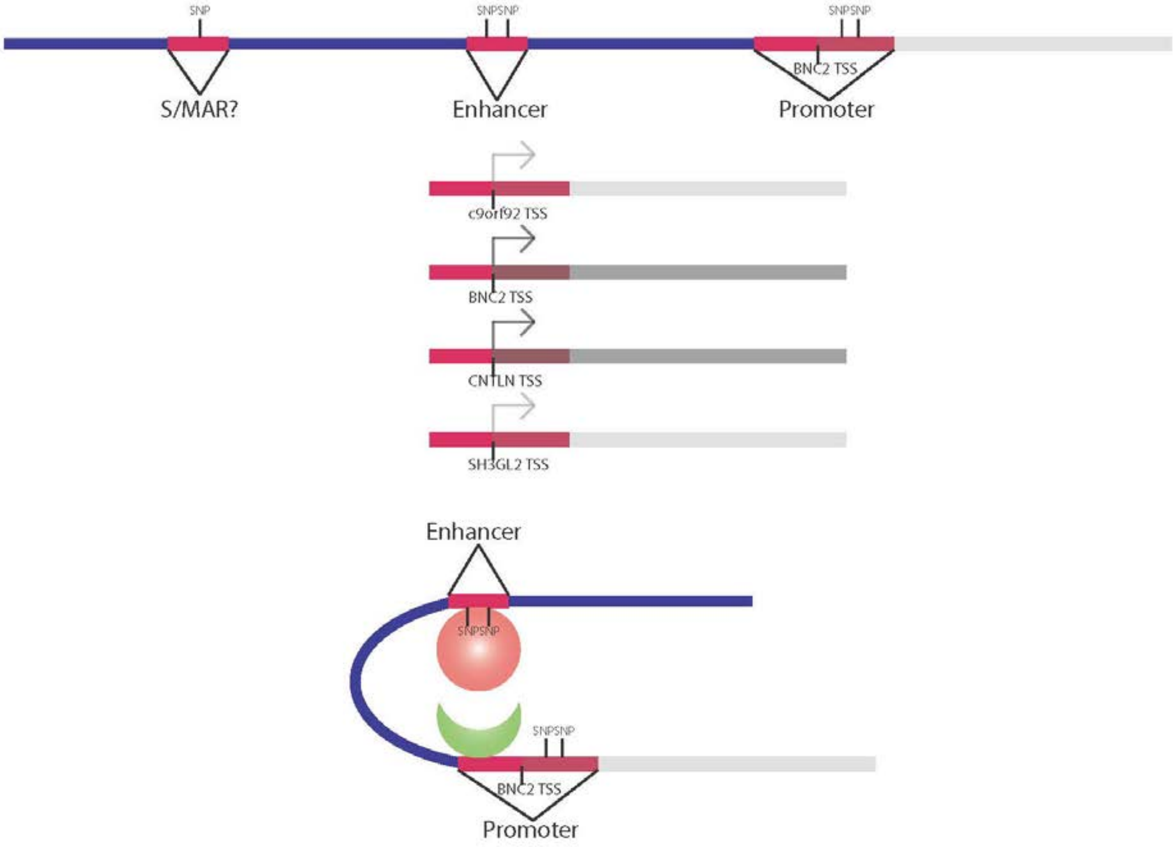


Figure 13. Summary of locus target genes. *BNC2* is the most likely target of the two SNPs in its promoter. The SNPs within the enhancer and S/MAR could possibly regulate genes within a 1 Mb region. Of the four genes, *CNTLN* and *BNC2* were the only ones expressed. The enhancer containing two functional SNPs looped to the promoter of *BNC2* and not the promoter of *CNTLN.*

Materials and Methods

eQTL Analysis

eQTL analysis was performed as described elsewhere (Pharoah et al., 2013).

3C

iOSE11 cells grown to approximately 80% confluence in media with a 1:1 ratio of MCDB105/Medium 199, 15% Fetal Bovine Serum (FBS), 10 ng/mL Epidermal Growth Factor, 0.5 µg/mL hydrocortisone, 5 µg/mL insulin, and 34 µg/mL Bovine Pituitary Extract were trypsinized and re-suspended in 1% formaldehyde. Fixed cells were pelleted and then re-suspended with 0.125 M glycine-PBS solution. Cells were lysed in 500 µL cold lysis buffer (10 mM Tris HCl pH 8.0, 10 mM NaCl, 0.2% NP40, Protease Inhibitor). After centrifugation the remaining pelleted nuclei was rinsed with 500µL New England Biolabs Buffer 2 (NEBuffer2), then re-suspended with 200 µL NEBuffer2. An additional 1320 µL NEBuffer2 was added along with 168 µL 1% SDS and incubated at 65°C for 12 minutes, followed by addition of 176 µL 10% Triton X-100. EcoR1 (375 units) was added in 150 µL NEBuffer2 and incubated for 24 hours at 37°C.

To stop digestion, 86 µL of 10% SDS was added and samples were incubated for 30 minutes at 65°C. Cells (4 x 10^7) were pooled and mixed with 7.44 mL of ligation buffer (1x T4 Ligase Buffer (NEB) 1% Triton-X 100, 1 mg/mL BSA) followed by the addition of 10 µL of T4 DNA ligase. Samples were incubated for 1 to 5 days at 16°C and then digested with proteinase K and de-cross-linked at 65°C overnight. DNA was then extracted with Phenol-Chloroform (Sigma) and ethanol precipitated. Re-hydrated DNA was then desalted with Microcon Ultra Cell YM -100 and eluted.

qPCR was performed by using Taq Polymerase PCR Kit (Qiagen) and Syto9 (Life Technologies) with 30 ng of DNA 1X Taq Buffer, 0.2 mM dNTPs, 0.25 µM Primers, 0.1 µL Taq Polymerase, 30 ng of DNA library, 5 µM Syto9; 95°C for 15 minutes, 50 cycles at 94°C for 20 seconds, 60°C for 1 minute. Samples were run using FAM Spectrum on and Applied Biosystems 7900 HT Fast Real Time PCR System. EcoR1 digested BACs (RPCI-11-185E1 Empire Genomics, RPCI-11-179K24 Life Technologies, RPCI-11-106G11 Life Technologies) for the region were used for the standard curve (Hagege et al., 2007). Interactions were calculated as a percentage of a restriction site directly adjacent to the bait restriction site. Sites with significantly higher frequency of interaction than the site adjacent to the anchor were considered significant. 3C was performed with two biological replicates and three technical replicates.

CHAPTER FOUR:

FUNCTIONAL ANALYSIS OF BNC2

Note to Reader

A manuscript that has been submitted for review includes portions of this chapter.

Review of *Basonculin 2*

Taken together, the functional data points to *BNC2* as the most likely candidate target gene of the causal SNPs at the ovarian cancer susceptibility locus.

Identification of *Basonuclin 2* and comparison to *Basonculin 1*

A chicken and mouse EST database search for homologs of *Basonuclin 1 (bnc1)* identified the novel gene *bnc2* (Romano et al., 2004; Vanhoutteghem and Djian, 2004). *Bnc1* expresses in epidermal keratinocytes and reproductive germ cells of mouse testis and ovary (Mahoney et al., 1998; Tseng and Green, 1992; Yang et al., 1997). BNC1, a nuclear ZF protein and TF, binds to and activates transcription of ribosomal RNA promoters while also localizing to areas typical of RNA polymerase I TFs (Iuchi and Green, 1999; Tian et al., 2001; Tseng et al., 1999). ZF domains have specific amino acids within the alpha helix of the domain that interact with specific nucleotides within the major groove of the DNA helix (Wolfe et al., 2000). *BNC1* and *BNC2* only have

43.4% identity between their sequences, yet, BNC1 and BNC2 have a similar structure with three separated pairs of two cysteine and two histidine (C2H2) ZFs and a nuclear localization signal (NLS) between the first and second pair of ZFs (Romano et al., 2004; Vanhoutteghem et al., 2011; Vanhoutteghem and Djian, 2004). ZF 1,2 has the most similarity between *BNC1* and *BNC2* (92.1%) and has the most conservation across species while ZF 3,4 and ZF 5,6 have less similarity between the two genes (Romano et al., 2004; Vanhoutteghem et al., 2011). A highly conserved region between the two proteins also lies in the N-terminus, yet the region has no obvious functional domain (Vanhoutteghem et al., 2011; Vanhoutteghem and Djian, 2004). *BNC2* has 6 exons, 5 introns and spans over 300 kb while *BNC1* has 5 exons and spans over 29 kb (Romano et al., 2004). Yet both genes are split up almost identically (Romano et al., 2004). Both genes have a GC rich TATA-less promoter (Vanhoutteghem et al., 2011; Vanhoutteghem and Djian, 2004) and therefore may have similar transcription regulation. Mammals, birds, and fish express both genes (Lang et al., 2009; Romano et al., 2004; Vanhoutteghem and Djian, 2004). Additional orthologs with a similar structure to the Basonuclins includes the *disco* genes in *Drosophila* and other insects (Romano et al., 2004). A single *Basonuclin* expresses in invertebrate chordates *Branchiostoma floridae* and *Ciona intestinalis* which resembles *BNC2* (Vanhoutteghem et al., 2011). Therefore a duplication to create two Basonuclins occurred in a vertebrate ancestor and *BNC2* is the more ancient gene and most likely more important (Vanhoutteghem et al., 2011). *BNC2* is extremely conserved and more so than *BNC1* with a sequence identity of 97.2% between human and mouse (Vanhoutteghem and Djian, 2004). Out of fifty

C2H2 ZF proteins, BNC2 is the ninth most conserved protein further indicating an essential function for *BNC2* (Vanhoutteghem and Djian, 2006).

Expression of *BNC2*

In the mouse, *bnc2* expresses highly in the skin tissue, ovary, and kidney with low levels in the testes, small intestine, and lung (Romano et al., 2004). No expression was seen in the peripheral blood leukocytes, brain, colon, liver, spleen or thymus (Romano et al., 2004; Vanhoutteghem and Djian, 2004). Closer inspection of expression in mice revealed expression in the mesenchymal cells, mainly connective tissue surrounding specific organs including the brain meninges, the cartilage, and the bone as well as the male germ cells (Vanhoutteghem et al., 2011; Vanhoutteghem et al., 2009; Vanhoutteghem et al., 2014). It is not expressed in the female germ cells (Vanhoutteghem et al., 2014). In zebrafish, *bnc2* expresses in the hypodermis, somatic cells of ovaries, brain, dorsal spinal cord, eye, superficial cells of vertebrae, fins, gut, kidney, and testes (Lang et al., 2009). *Drosophila* and insect *disco* proteins have similar expression to *BNC2* (Vanhoutteghem et al., 2009).

Interestingly transcription of the human *BNC2* gene can be initiated at 6 different exons with the canonical form being the most abundant (Vanhoutteghem and Djian, 2007). The human *BNC2* gene also has alternative splice sites in the downstream exons adding up to a total of 23 different exons (Vanhoutteghem and Djian, 2007). There is potential for alternative transcripts in mice but not nearly as many in humans indicating that the non-conserved transcripts play less of an important role (Vanhoutteghem and Djian, 2007). There also seems to be tissue specific splicing since reverse-transcriptase PCR revealed different isoforms among human testes, kidney, and keratinocytes (Vanhoutteghem and Djian, 2007). As mentioned in the previous paragraph, *BNC2* is expressed in several different cell types and therefore may have many different transcripts among different cells. The Vanhoutteghem, et. al, 2007 study states in the title that “The human *basonuclin 2* gene has the potential to generate nearly 90,000 mRNA isoforms encoding over 2,000 different protiens” (Vanhoutteghem and Djian, 2007), a most astonishing amount. All identified transcripts in the Vanhoutteghem, el. al, 2007 study were transfected into HeLa cells and only four actually expressed (Vanhoutteghem and Djian, 2007). These four included the canonical form, one with a modified fourth finger (adds more residues between the second cysteine and first histidine), one that lacked ZF 5,6 and one that lacked all the ZFs and nuclear localization signal (Vanhoutteghem and Djian, 2007). Northern blot analysis in mouse shows the prescence of 9, 6, 4, and 2 kb transcripts (Romano et al., 2004). The isoforms appearing in both mouse and human suggests that these four isoforms may all play an important role in the tissues in which they are expressed.

Function of BNC2

As mentioned previously, BNC1, a nuclear ZF protein and TF, binds to and activates transcription of ribosomal RNA promoters while also localizing to areas typical of RNA polymerase I TFs (Iuchi and Green, 1999; Tian et al., 2001; Tseng et al., 1999). Romano et al., 2004 tested the ability of BNC2 ZFs to bind to the ribosomal RNA promoter in in vitro DNA binding assays and indeed BNC2 binds to the ribosomal RNA promoter (Romano et al., 2004). Unlike BNC1, BNC2 does not shuttle out of the nucleus

(Vanhoutteghem and Djian, 2006). BNC1 has a serine in the NLS that becomes phosphorylated in order to transport into the cytoplasm (Vanhoutteghem et al., 2011; Vanhoutteghem and Djian, 2006). BNC2 has a proline at this residue and therefore cannot be phosphorylated (Vanhoutteghem et al., 2011; Vanhoutteghem and Djian, 2006). Interestingly, BNC2 was first seen as a 145 kDa protein to come down in the insoluble fraction of the nuclear extract indicative of localizing to nuclear speckles (Vanhoutteghem and Djian, 2006). Proteins involved in splicing, mRNA export, nonsense mediated decay and polyadenylation (Fu and Maniatis, 1990; Kataoka et al., 2000; Krause et al., 1994; Lamond and Spector, 2003; Zhou et al., 2000) also localize to nuclear speckles potentially revealing a function for BNC2. Vanhoutteghem et al. also saw that BNC2 co-localized with SC35, a splicing factor (Vanhoutteghem and Djian, 2006). Interestingly, the Phillipe Djian group also identified a larger isoform (160 kDa) of BNC2 that comes down in the soluble fraction of the nuclear extract and appears in the chromatin fraction indicative of BNC2 acting as a typical TF (Vanhoutteghem et al., 2011; Vanhoutteghem et al., 2014). This isoform is actually the most abundant and canonical isoform in mice (Vanhoutteghem et al., 2014). The isoform that locates to nuclear speckles does not appear in mice (Vanhoutteghem et al., 2014).

Animal models with disruption of *bnc2* gave intriguing phenotypes. Zebrafish with truncating mutations of *bnc2* lack body stripes, are significantly smaller than wild type, and females are infertile (Lang et al., 2009). The dysmorphic ovaries in zebrafish provide the first link between ovarian development and *bnc2*. The Lang et al. paper noted that the oocytes from the mutants were fertilizable and that the infertility most likely stems from the excess somatic ovarian tissue (Lang et al., 2009). This group also discovered that bnc2 acts non cell autonomously on the zebrafish stripes (the iridescent iridiphores, the orange/yellow xanthaphores, and the black melanophores) (Lang et al., 2009). *Bnc2* is not expressed in the pigmented cells but resides in the hypodermis (Lang et al., 2009). *Bnc2* mutant mice are also dwarfed in size compared to wild-type and infertile like the zebrafish mutants (Vanhoutteghem et al., 2011; Vanhoutteghem et al., 2014). Interestingly, *bnc2* mutant mice die neonatally due to cleft palate and other craniofacial and tongue abnormalities (Vanhoutteghem et al., 2011; Vanhoutteghem et al., 2009). *Bnc2* is strongly expressed in the craniofacial sites affected; therefore bnc2 has a direct effect on the affected cells (Vanhoutteghem et al., 2009) unlike the indirect effect on the pigmented cells of zebrafish. Vanhoutteghem et al. also discovered that the craniofacial abnormalities were due to less cells entering mitosis during development (Vanhoutteghem et al., 2009). Due to the expression in male germ cells, Vanhoutteghem et al. investigated its role in spermatogenesis and discovered that *bnc2* expression regulates mitosis of prospermatagonia and prevents meiosis in fetal testis (Vanhoutteghem et al., 2014). Male germ cells undergo meiosis after birth while female germ cells undergo meiosis in the fetus therefore could explain why *bnc2* is expressed in male but not female germ cells (Vanhoutteghem et al., 2014).

The 9p22.2 locus containing the *BNC2* gene has also been identified as a human height and skin pigmentation GWAS locus (Eriksson et al., 2010; Hider et al., 2013; Jacobs et al., 2013; Visser et al., 2014; Wood et al., 2014). These studies implicate *BNC2* as the target gene of this locus since traits affected by BNC2 mutation or knock out in animal models affect overlapping traits associated with 9p22 in human GWAS. Further understanding of how this gene functions may reveal how changes in the transcription rates of *BNC2* lead to ovarian cancer predisposition and development. Here we characterize BNC2 as a TF and found that this protein acts primarily by binding a specific DNA sequence located in enhancer regions proximal to genes involved in cell-cell communication and chromatin remodeling. This work reveals a complex regulatory network in a cancer risk locus and provides new insights into the etiology of EOC.

Results

C2H2 Zinc Finger Proteins

The most obvious indicator of function for BNC2 is the presence of three separated pairs of C2H2 ZF domains (Romano et al., 2004; Vanhoutteghem and Djian, 2004). A thorough understanding of how these domains work will give clues for the role of *BNC2* in ovarian cancer susceptibility. These domains contain two cysteine and two histidine residues that, together, bind to zinc stabilizing a fold which creates two beta sheets and an alpha helix (Wolfe et al., 2000). ZF domain containing proteins comprise half of all known human and mouse TFs (Emerson and Thomas, 2009; Fulton et al., 2009; Messina et al., 2004; Tupler et al., 2001) and are the largest protein family in mammalian genomes (Ravasi et al., 2003). Approximately 700 human C2H2 ZF proteins have been identified (Najafabadi et al., 2015; Vaquerizas et al., 2009; Weirauch and Hughes, 2011).

C2H2 ZFs domains display sequence specific DNA binding through amino acids at positions -1, 2, 3, and 6 of the ZF alpha helix with each DNA binding amino acid binding to one nucleotide (Elrod-Erickson et al., 1996; Klug, 2010; Lam et al., 2011; Pavletich and Pabo, 1991; Wolfe et al., 2000). DNA binding usually requires two or more ZF domains in tandem and these tandem ZF domains overlap binding with the previous ZF binding site at the fourth base pair (Elrod-Erickson et al., 1996; Pavletich and Pabo, 1991; Wolfe et al., 2000). The position 6 amino acid binds to the first nucleotide on the primary strand (which could also be the fourth nucleotide bound to a ZF in tandem), position 3 binds to the second, and -1 to the third. The position 2 amino acid binds to the fourth nucleotide on the complementary strand (Wolfe et al., 2000).

Due to the unique way these ZF domains bind to DNA, many groups have attempted to identify a “recognition code” for ZF domains and possibly design ZFs for gene therapy, yet specificity has not been optimal (Corbi et al., 1998; Corbi et al., 1997; Wolfe et al., 2000). Difficulty obtaining the “recognition code” most likely stems from adjacent ZFs overlapping with the DNA sequence they recognize altering the individual ZFs ability to bind to the predicted sequence (Isalan et al., 1997). Also, interactions between amino acids of neighboring ZFs and within individual ZFs (outside of amino acids at positions   -1, 2, 3, and 6) can orient the protein in a way that alters the sequence it should typically recognize (Wolfe et al., 2001). The specificity of these proteins has been questioned but Lam et al. clearly explained that C2H2 ZFs in tandem bind to degenerate motifs, in other words, they bind to many different but similar motifs (Lam et al., 2011).

Najafabadi et al. developed an improved recognition code by testing 47,072 natural ZF domains sampled from all eukaryotes in tandem against a library of possible nucleotide binding sequences using a bacterial one hybrid system (Najafabadi et al., 2015). This data was used to create a model that takes into account the degeneracy of the binding (Najafabadi et al., 2015). Many bacterial one hybrid sequences and predicted sequences were confirmed via ChIP-seq and protein binding microarray (PBM) (Najafabadi et al., 2015). A PBM is an array with double stranded DNA probes representing all possible 8mer sequences DNA binding proteins can recognize (Berger and Bulyk, 2009; Berger et al., 2006). Glutathione-S-transferase (GST)-tagged DNA binding proteins are applied to the array followed by a fluorophore for GST to asses which sequences the protein recognizes (Berger and Bulyk, 2009; Berger et al., 2006). It has been shown that C2H2 ZF domains can also bind RNA or proteins, but Najafabadi et al. clearly demonstrates that the majority bind DNA (Najafabadi et al., 2015). Again, due to the unique nature in which ZF domains recognize and bind to nucleotide motifs, these TFs have the most unique binding sites compared to any other TFs (Najafabadi et al., 2015).

BNC2 Zinc Fingers Recognize Specific DNA Sequences *in vitro*

BNC2 contains three pairs of C2H2 ZFs suggesting that it interacts with specific DNA sequences and plays a role in transcription regulation (Figure 15A) (Vanhoutteghem and Djian, 2004, 2006). In order to identify potential DNA sequences recognized by the BNC2 ZFs, bacterially expressed GST-tagged constructs of each ZF pair (Figure 14A) were applied to a protein binding microarray (PBM) of overlapping, rationally randomized nucleotides (Berger and Bulyk, 2009; Berger et al., 2006; Lam et al., 2011). The top ten scoring sequences for each of the individual ZF pairs were then aligned to generate a logo using position weight matrix scoring (Figure 15A). The motifs

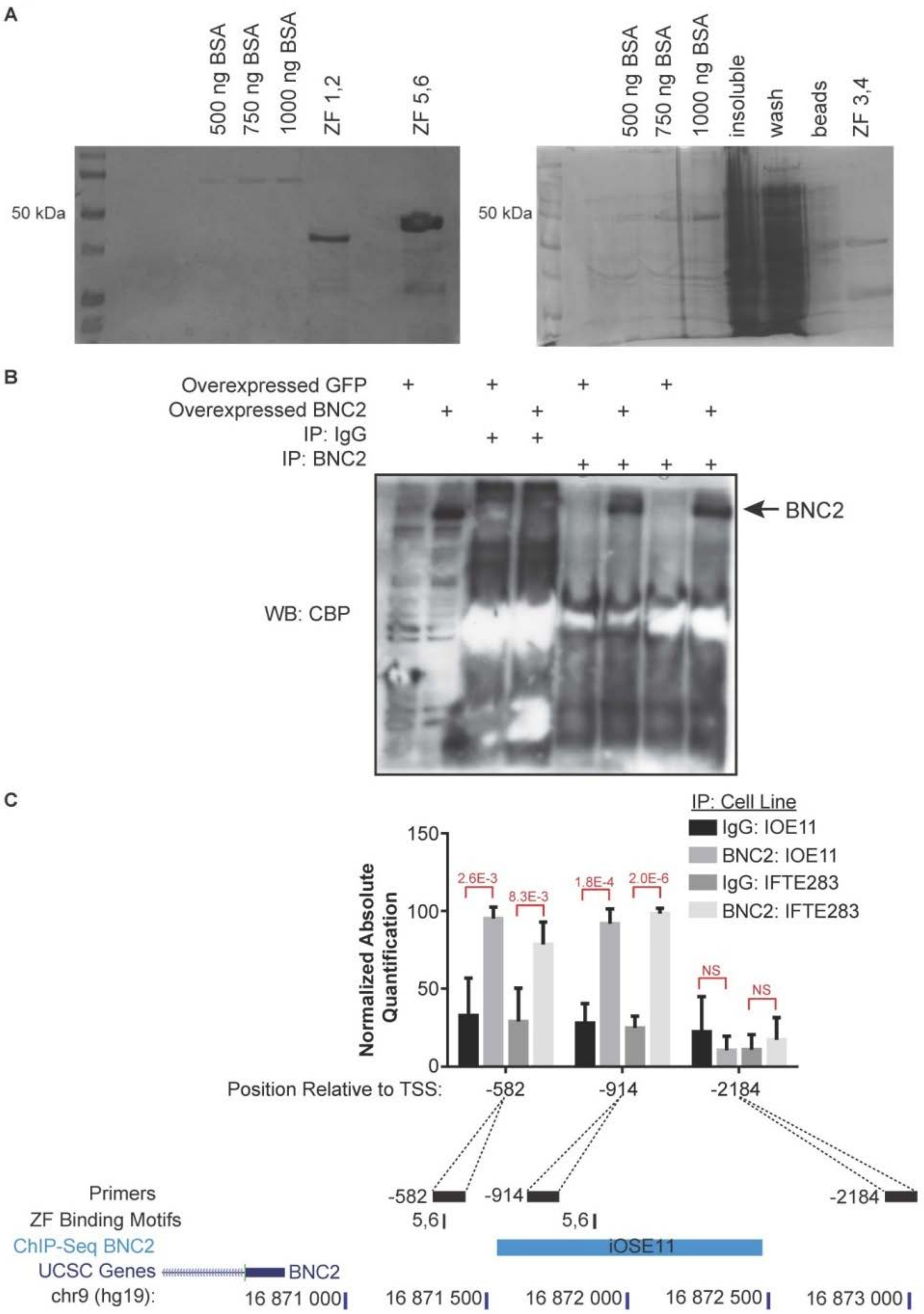


Figure 14: BNC2 binds to its own promoter. A. Coomassie stain of protein purification of GST tagged BNC2 ZF pairs: 1,2; 3,4; and 5,6. **B.** CPB tagged GFP and BNC2 were over expressed in 293FT cells. Lysates of these cells were

immunoprecipitated with either Rabbit IgG or the Prestige antibody for BNC2 (Sigma). Immunoprecipitates undergo Western Blot for CBP. A band for BNC2 between the 150 kDA and 250 kDA mark appears in the input and BNC2 IP for over expressed BNC2 but not in the input and BNC2 IP for over expressed GFP nor in the IgG IP. **C.** ChIP indicates that BNC2 binds to its own promoter. Potential ZF 5,6 binding sites within the *BNC2* promoter are indicated with black lines. Black boxes indicate location of amplicons analyzed with ChIP qPCR. In iOSE11 and iFTSEC283 cells there is a signal that BNC2 is indeed binding to those sites (bar graph). Raw ChIP-seq data for BNC2 in iOSE11 cells replicate the binding at the -914 position (blue bar).

for ZF1,2 and 5,6 are almost identical to the predicted C2H2 “recognition code” (Najafabadi et al., 2015). The data for ZF3,4 yielded lower-confidence data and did not match the recognition code predictions (Figure 15A) (Berger et al., 2006). The 3’ end of the ZF1,2 and ZF5,6 binding motifs contain the same nucleotides at the same position and weight, consistent with the similarity between ZF2 and ZF6 in amino acid residues at positions that specifically interact with DNA (Figure 15A). Interestingly, the *BNC2* promoter contains two of the top 10 BNC2 ZF5,6 PBM binding sequences (Figure 14C).

To validate BNC2 binding sequences identified using the PBM we conducted ChIP in iOSE11 and iFTSEC283 cells for endogenous BNC2 (Figure 14B) to determine its presence at the putative PBM sites (-582 and -914 base pair (bp) upstream of the TSS) found at the *BNC2* locus as well as at site >300 bp (-2184) from the TSS as a negative control. A significantly larger amount of DNA pulled down with the BNC2 antibody than with the IgG control at the -582 (iOSE11 p = 2.6 x 10^{-3}, iFTSEC283 p = 8.3 x 10^{-3}) and -914 (iOSE11 p = 1.8 x 10^{-4}, iFTSEC283 p = 2.0 x 10^{-6}) bp sites, but not at the -2184 bp site (Figure 14C). This supports that BNC2 recognizes the sites identified in the PBM experiment and also suggests an auto-regulatory mechanism.

BNC2 Genome-wide Target Sites

To determine the sites in the genome bound by BNC2 in ovarian cells we conducted ChIP-seq in iOSE11 and iFTSEC283 cells. MEME, a motif analysis tool, identified a motif centrally enriched in the ChIP-seq peaks for both cell types (Figure 15B). ChIP-seq data replicated BNC2 binding in the iOSE11 cells at the -914 position tested in ChIP-qPCR (Figure 14C). Interestingly the motif identified by MEME seems to be a concatenation of the motif for ZF1,2 and the reverse complement motif for ZF5,6 (Figure 15B). ZF1,2 and 5,6 are greater than 500 amino acids away from each other potentially allowing the protein to fold in a way that allows the two ZF paired domains to bind to the DNA as dimers (Figure 15B). The originally identified sequence has a 75% homology to the MEME motif (Figure 15B). About 50% and 25% of the iFTSEC283 and iOSE11 peaks, respectively, have the motif near the peak summits (Figure 15C).

We annotated the transcriptional landscape of the BNC2 ChIP-seq peaks in iOSE11 cells by overlapping them with ChIP-seq for H3K27ac and H3K4me1 in iOSE11 cells and at core promoters (1 kb of TSS) (Figure 15D). BNC2 ChIP-seq peaks that overlap with H3K4me1 and H3K27ac were considered regions of active enhancers. Peaks that only overlap with H3K27ac were considered active chromatin. Peaks that only overlap with H3K4me1 were considered poised enhancers. Sixty-six percent of BNC2 ChIP-seq peaks overlap with a regulatory element. Interestingly, a small percentage of BNC2 recognition sites reside in core promoters indicating *BNC2* works, in part, by modulating the activity of enhancers.

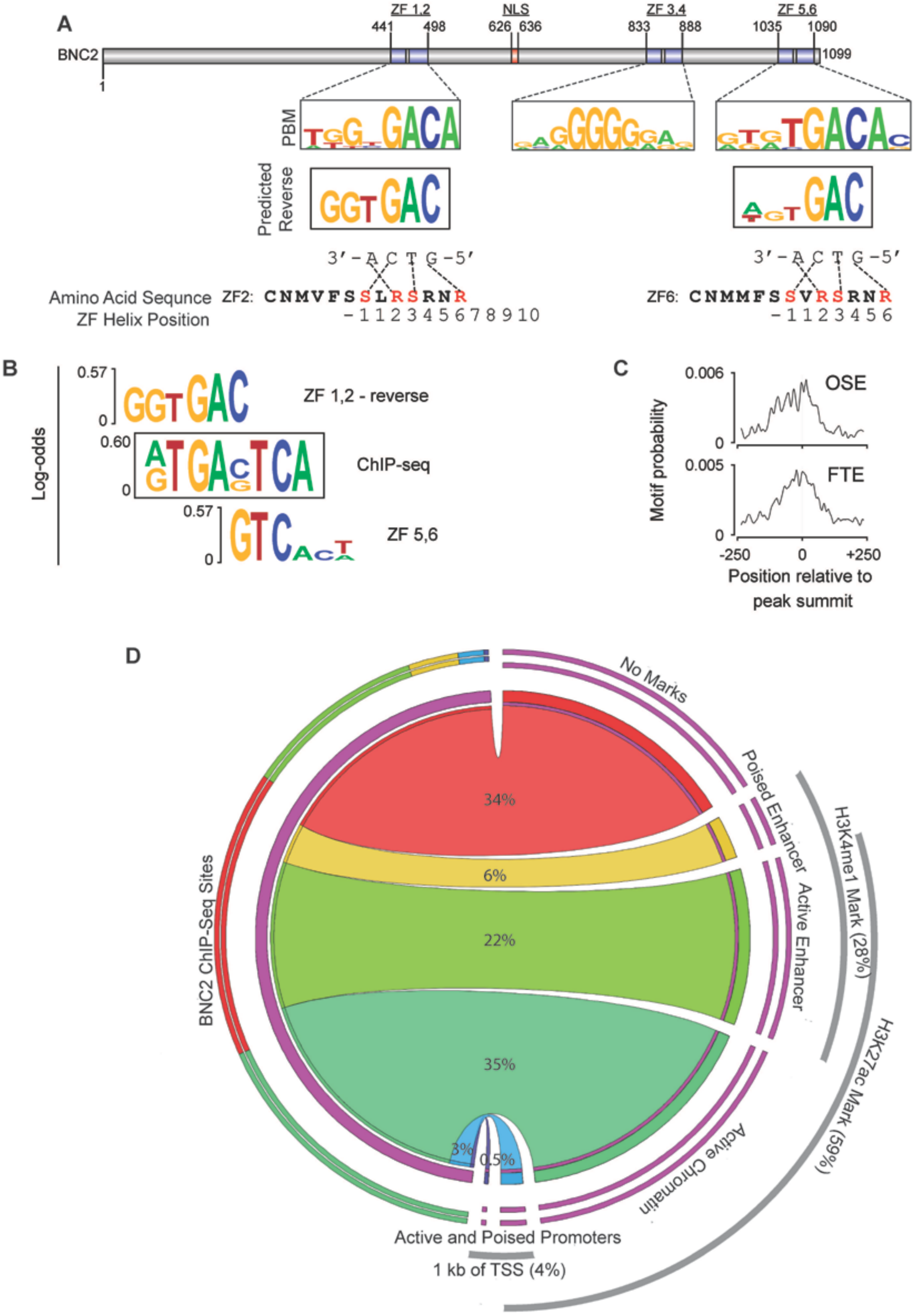
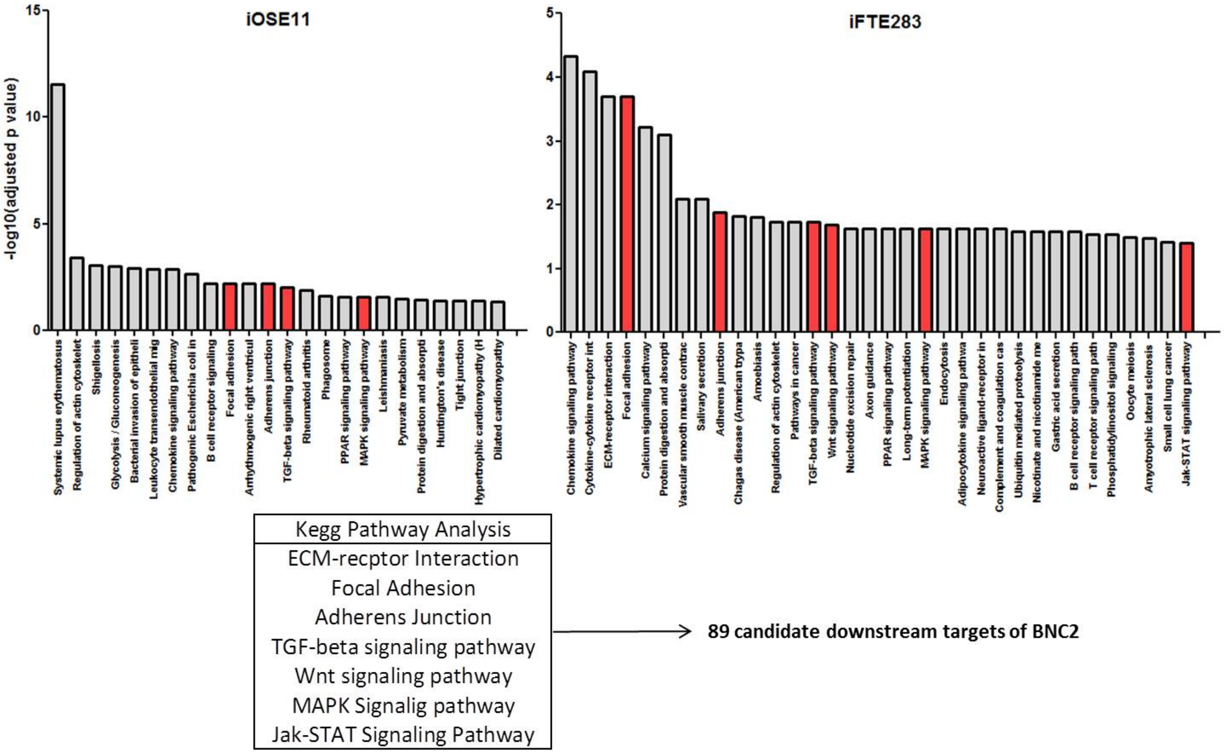


Figure 15. BNC2 recognizes specific nucleotide sequence. *A.* BNC2 is characterized as a C2H2 ZF protein with three pairs of ZFs (called 1,2; 3,4; 5,6). BNC2 ZF binding sites were identified in vitro by applying recombinant proteins of each ZF pair

to a PBM. Position weight matrices of all potential binding sites with significant scores for each BNC2 ZF pair are shown as logos. Motifs predicted based on the protein sequence of the ZF domains aligned with ZF1,2 and ZF5,6. The 3' end of the sequences recognized by ZF1,2 and ZF5,6 reveal the same nucleotides. Inspection of the amino acid sequences for ZF2 and ZF6 show that amino acid residues at position -1, 2, 3, and 6 within the alpha helix that specifically interact with DNA nucleotides (in red) are the same. **B.** The ChIP-Seq motif identified by MEME seems to be a concatenation of the predicted motif for ZF1,2 and the predicted reverse complement motif for ZF 5,6 or vise-a-versa. **C.** Enrichment of motif relative to ChIP-Seq peak summits. **D.** A Circos table depicts the percentage of BNC2 ChIP-seq sites overlapping with chromatin mark's ChIP-seq sites in iOSE11 cells. Sites containing only H3K4me1 marks are considered poised enhancers. Sites containing H3K4me1 and H3K27ac marks are considered active enhancers. Sites containing only H3K27ac marks are considered active chromatin. Sites within 1 kb of a TSS that contain H3K27ac marks are considered active promoters. Sites within 1 kb of a TSS without histone marks are considered poised promoters.

Identification and Validation of BNC2 Target Genes

To identify target genes regulated by enhancers containing BNC2 binding sites we used the Galaxy Cistrome program (Liu et al., 2011). This generated a list of 445 genes in iOSE11 cells and 725 genes in iFTSEC283 cells with TSS within 30 kb of the BNC2 ChIP-seq peak centers. One hundred and sixty eight genes lie near BNC2 ChIP-seq peaks in both cell types. KEGG Pathway analysis identified several pathways that are likely targets including chemokine and TGF-beta signaling pathways (Figure 16). Next, we selected a set of 89 genes implicated in ovarian cancer, ovarian cancer GWAS, follicular development, ovarian development (Bojesen et al., 2013; Cancer Genome Atlas Research, 2011; Goode et al., 2010; Nef et al., 2005; Permuth-Wey et al., 2013; Pharoah et al., 2013; Ramakrishna et al., 2010; Schindler et al., 2010; Shen et al., 2013; Song et al., 2009), were part of the significant KEGG pathways, or contained ChIP-seq peaks within their core promoter (within 1kb from the TSS)(Table 3)



Kegg Pathway Analysis

ECM-recptor Interaction
Focal Adhesion
Adherens Junction
TGF-beta signaling pathway
Wnt signaling pathway
MAPK Signalig pathway
Jak-STAT Signaling Pathway

89 candidate downstream targets of BNC2

Figure 16: Kegg Pathway analysis of downstream targets of BNC2. A Kegg Pathway analysis on all genes within 30 kb of BNC2 ChIP-seq sites in iOSE11 and iFTE283 cells revealed several pathways reaching significance. Bar graphs in red indicate the pathways relevant to cancer. Genes within 30 kb of BNC2 ChIP-seq sites and within those pathways were analyzed further.

and tested whether overexpression of *BNC2* in HEK 293T (Figure 17) would modulate their expression measured by Nanostring. Several genes implicated in ovarian cancer and ovarian development or that mapped to KEGG Focal Adhesion, ECM-receptor interaction or TGF-β Signaling Pathways showed changes in expression induced by *BNC2* overexpression as measured 24h after transfection (Table 4).

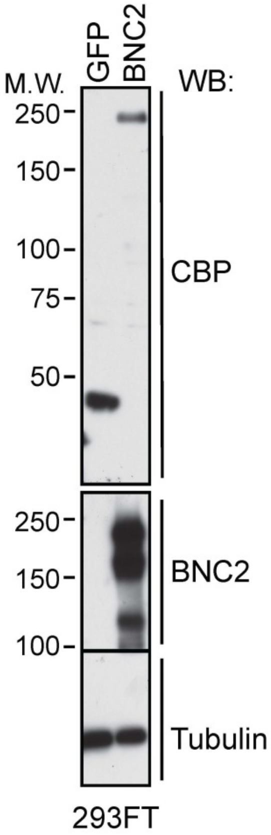


Figure 17. Overexpression of *BNC2* in HEK293FT cells. Western blot indicates the over-expression of CBP-tagged BNC2 in 293FT cells. Blotting for the CBP tag and the antibody for BNC2 clearly show that *BNC2* is over-expressed.

Table 3: Expression analysis of downstream target genes of BNC2

Gene	Cell Line	Pathway	P-Value
FAM49B	FTE	Ovarian Cancer	0.00015
ITGB5	FTE	ECM-receptor interaction, Focal Adhesion	0.00046
PKDCC	FTE	Ovarian Development	0.00066
CCND3	FTE	Focal adhesion, WNT Signaling Pathway, JAK-STAT Signaling Pathway	0.00107
CEP55	FTE	Ovarian Development	0.00264
GUSB		Reference	0.00528
JUN	FTE, OSE	Focal adhesion, WNT Signaling Pathway, MAPK Signaling Pathway	0.00627
COL6A3	FTE	ECM-receptor interaction, Focal Adhesion	0.00698
ITGA3	OSE	Focal Adhesion	0.00711
CAPN2	FTE	Focal adhesion	0.00943
SLFN12	FTE	Promoter with Peak	0.01012
FBXO15	OSE	Promoter with Peak	0.01375

Table 3 (Continued)

Gene	Cell Line	Pathway	P-Value
COL4A5	FTE	ECM-receptor interaction, Focal Adhesion	0.01748
FEM1A	OSE	Promoter with Peak	0.02064
TTI2		Reference	0.02961
TGFBR3	FTE	TGF-beta Signaling Pathway	0.03906
STK35		Reference	0.0451
BMP6	FTE	TGF-beta Signaling Pathway	0.05324
SMG5		Reference	0.05469
SNAI2	FTE	Adherens Junction	0.06178
PPP3CA	FTE	WNT Signaling Pathway, MAPK Signaling Pathway	0.06363
FN1	FTE	ECM-receptor interaction, Focal Adhesion	0.06388
FASLG	FTE	MAPK Signaling Pathway	0.07104
RANBP10	OSE	Promoter with Peak	0.07547
BANP	OSE	Promoter with Peak	0.0801
INO80		Reference	0.0833
SH2D4A	FTE, OSE	Ovarian Development	0.08539
TOX4		Reference	0.09168
MAP3K8	FTE	MAPK Signaling Pathway	0.09761
ID1	OSE	TGF-beta Signaling Pathway	0.10798
LSM5	FTE	Follicular Development	0.10818
ACTG1	OSE	Focal Adhesion, Adherens Junction	0.114
GBA	OSE	Promoter with Peak	0.1165
PPP6R3	OSE	Promoter with Peak	0.11658
POLDIP3	OSE	Promoter with Peak	0.11742
CD59	FTE	Promoter with Peak	0.12042
ANK3	FTE	Ovarian Development	0.13177
THOC6	Ose	Promoter with Peak	0.13242
LEP	FTE	JAK-STAT Signaling Pathway	0.14078
HNF1B	FTE	GWAS	0.14732
LMO7	FTE	Adherens Junction	0.1582
DKK1	FTE	WNT Signaling Pathway	0.16114
ACTB		Reference	0.17339
MYLK	FTE	Focal adhesion	0.17567
LEPR	FTE	JAK-STAT Signaling Pathway	0.17834
ACTN4	OSE	Focal Adhesion, Adherens Junction	0.21508
KBTBD6	OSE	Promoter with Peak	0.21679
FSCN1	OSE	Promoter with Peak	0.23926
BTBD19	OSE	Promoter with Peak	0.25273
LAMB1	FTE	ECM-receptor interaction, Focal Adhesion	0.25369
TCTEX1D4	OSE	Promoter with Peak	0.26229
THBS1	FTE, OSE	ECM-receptor interaction, Focal Adhesion, TGF-beta Signaling Pathway	0.27286

Table 3 (Continued)

Gene	Cell Line	Pathway	P-Value
CD36	FTE	ECM-receptor interaction	0.27883
EIF2C1		Reference	0.29711
SPHK1	OSE	Ovarian Development	0.30676
ITGA5	OSE	Focal Adhesion	0.31715
RASGRP1	FTE	MAPK Signaling Pathway	0.33703
FAP	FTE	Ovarian Cancer	0.33708
RASGRP3	FTE, OSE	MAPK Signaling Pathway	0.34904
RASGRF2	OSE	MAPK Signaling Pathway	0.35168
WASL	FTE, OSE	Adherens Junction	0.35776
DUSP1	OSE	MAPK Signaling Pathway	0.36196
PDCD1LG2	FTE	Promoter with Peak	0.37688
CCDC80	FTE	Promoter with Peak	0.39362
PRDM4		Reference	0.39529
SMAD3	FTE, OSE	Adherens Junction, TGF-beta Signaling Pathway, WNT Signaling Pathway	0.39745
CTGF	FTE, OSE	Follicular Development	0.40642
TRIB2	FTE	Ovarian Development	0.43142
SPARC	FTE	Follicular Development	0.44907
PDGFC	FTE	Focal adhesion	0.47425
CLPB	OSE	Promoter with Peak	0.49331
COL4A6	FTE	ECM-receptor interaction, Focal Adhesion	0.49815
CACNG6	OSE	MAPK Signaling Pathway	0.51917
UHRF1BP1L	FTE	Ovarian Development	0.54734
LTBP1	FTE	TGF-beta Signaling Pathway	0.56817
IL22	FTE	JAK-STAT Signaling Pathway	0.57974
IFNGR1	FTE	JAK-STAT Signaling Pathway	0.58208
IL22RA2	FTE, OSE	JAK-STAT Signaling Pathway	0.58664
CD44	FTE	Ovarian Development, ECM-receptor interaction	0.5942
RAI14	FTE	Promoter with Peak	0.63357
DUSP3	FTE	MAPK Signaling Pathway	0.64444
PRKCA	FTE	Focal adhesion, WNT Signaling Pathway, MAPK Signaling Pathway	0.66141
RPS6KA2	FTE	MAPK Signaling Pathway	0.6638
DUSP10	FTE, OSE	MAPK Signaling Pathway	0.67452
GNA12	OSE	MAPK Signaling Pathway	0.70082
TSNAXIP1	OSE	Promoter with Peak	0.71591
BCL2L1	FTE	JAK-STAT Signaling Pathway	0.73164
AP2S1	OSE	Promoter with Peak	0.74506
TGFBR2	FTE	Adherens Junction, TGF-beta Signaling Pathway, MAPK Signaling Pathway	0.76959
ITGB2	OSE	promoter with Peak	0.83732
SMAD7	FTE, OSE	TGF-beta Signaling Pathway, Promoter with Peak	0.85264
PLOD2	FTE	Ovarian Development	0.85924

Table 3 (Continued)

Gene	Cell Line	Pathway	P-Value
GIGYF2		Reference	0.8774
GADD45B	OSE	MAPK Signaling Pathway	0.89367
ZBTB4	OSE	Promoter with Peak	0.89726
PHF12	OSE	Promoter with Peak	0.89743
SEMA3C	OSE	Ovarian Development	0.90991

BNC2 Interacts with the NuRD Complex

To further elucidate the function of BNC2 in transcription regulation we used tandem affinity purification to isolate proteins that are in complex with ectopically expressed tagged *BNC2* in HEK293T cells (Figure 18A). Gene ontology enrichment analysis using the web gestalt program (http://bioinfo.vanderbilt.edu/webgestalt/) (Zhang et al., 2005)

indicates that BNC2 interacts with proteins enriched in the NuRD complex of transcription repression (p = 2.75 x 10^{-11}) (Figure 18B, red circle), RNA binding proteins (p = 9.47 x 10^{-6}), and proteins involved in gene expression (p = 2.78 x 10^{-6}). Immunoprecipitation of endogenous MTA2, a component of the NuRD complex, followed by a western blot for endogenous BNC2 was used to verify the interaction (Figure 18C).

Since BNC2 interacts with the NuRD complex we then tested the extent to which BNC2 constructs fused to GAL-4 DNA binding domain represses transcription of a heterologous promoter (Figure 18D). Full length and fragments of BNC2 significantly repressed expression of luciferase compared to GAL-4 alone indicating that BNC2 displays transcription repression activity and may interact with the NuRD complex via multiple regions (Figure 18E).

Table 4: Genes with significant expression changes upon differential expression of *BNC2*.

iFTE283

Gene	TSS to Peak Center	P-Value	Expression Correlation to *BNC2*	Pathway
FAM49B	21553	0.00014782	-	Ovarian Cancer
PKDCC	16237	0.00066222	-	Ovarian Development
COL4A5	24983	0.0174824	-	Focal Adhesion
CAPN2	5635|-5189	0.00942743	-	Focal adhesion
ITGB5	8074	0.00046358	-	Focal Adhesion
COL6A3	-17808	0.00698256	+	Focal Adhesion
JUN	20445	0.00626649	+	Focal adhesion
TGFBR3	5215|-14508	0.0390608	+	TGF-beta Signaling Pathway
CCND3	-25565	0.00106617	+	Focal adhesion
SLFN12	-819	0.0101215	+	Promoter with Peak
CEP55	-24627|24647	0.002645	+	Ovarian Development

iOSE11

Gene	TSS to Peak Center	P-Value	Expression Correlation to *BNC2*	Pathway
JUN	-20156	0.00626649	+	Focal adhesion
ITGA3	-4619	0.00711417	+	Focal Adhesion
FBXO15	-620|-721	0.01375	+	Promoter with Peak
FEM1A	-65	0.0206431	-	Promoter with Peak

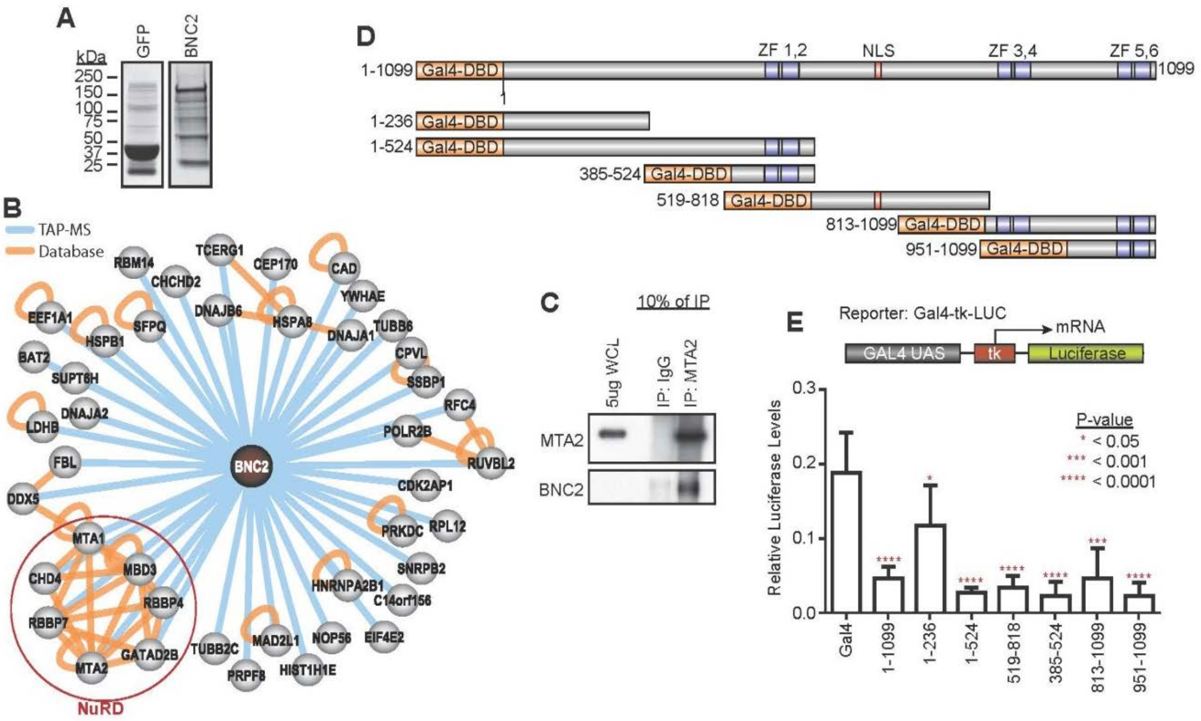


Figure 18: BNC2 interacts with the NuRD complex. A. TAP-tagged control GFP and BNC2 purified protein complexes from HEK293FT cells. ***B.*** Gray nodes represent proteins found in complex with BNC2. Blue and orange edges indicate interactions identified by tandem affinity purification coupled to mass-spectrometry or present in published datasets identified by the Cytoscape plugin BisoGenet, respectively. Notably, BNC2 interacts with proteins that are part of the NuRD complex (red circle). ***C.*** Confirmation immunoprecipitation indicates endogenous BNC2 interacts with NuRD protein MTA2. **D.** BNC2 constructs (gray boxes) fused to the GAL4 DNA binding domain (DBD) (orange boxes) were co-transfected with the GAL4-TK-Luc reporter vector. ***E.*** Illustration of the GAL4-tk-Luc construct (top). Measurements of relative luciferase levels determine which constructs have repression activity compared to the GAL4-DBD alone (bottom).

Summary

Here we have shown that BNC2 does indeed act as a TF. It recognizes a specific DNA binding motif in vitro which also appears as the predicted motif based on previous TF binding data. The motifs for ZF1,2 and ZF 5,6 also appear in the MEME analysis as a concatenation indicating that BNC2 binds to DNA in a folded form. BNC2 binding also overlaps with previously identified regulatory regions, more specifically enhancers, in ovarian cells. Changes in *BNC2* expression also lead to changes in downstream target gene expression. Interestingly many affected genes regulate cell-cell communication. A previous study suggests that BNC2 acts non-cell autonomously in the development of the Xanthaphores (stripes) in Zebrafish. Therefore, BNC2 may play a similar role in ovarian development. BNC2 also interacts with other TFs. Most interestingly, BNC2 interacts with the NuRD complex, a complex involved in chromatin remodeling and histone deacetylation. In general the NuRD complex acts as a repressor of transcription and a reporter assay with BNC2 constructs suggests that BNC2 represses transcription in this context.

Materials and Methods

Protein Binding Microarray

Fragments containing cDNAs of the each ZF pairs were PCR amplified from a plasmid containing full length *BNC2* cDNA (gift from Dr. Philippe Djian). Primers containing Gateway recombination sites are described in Table 5. PCR products were cloned into pDONR221 using the BP recombination kit and transferred to pDEST15 as a fusion to GST using LR recombination kit (Invitrogen). After plasmid purification, pDONR221 plus ZF plasmids undergo into using the LR recombination kit from Invitrogen. Plasmids containing GST-tagged ZFs were expressed in BL21 *E. coli* and purified using GT sepharose beads. Purified GST-ZFs were eluted from beads with 50 mM reduced glutathione (Figure 14A). The eluate was then dialyzed in TBS with 50 µM

Zinc Acetate and proteins were quantified using Bio-Rad Protein Assay. For the Protein Binding Microarray, 0.5 µg of each GST-ZF protein construct were applied individually to two differently designed arrays designated ME and HK as previously described (Berger and Bulyk, 2009; Lam et al., 2011). ZFs typically bind to degenerate motifs and have the potential to have more than one recognition sequence (Lam et al., 2011; Ramirez et al., 2008). Each DNA probe sequence is given an E-score which is similar to the Area under the ROC curve statistical metric and an E-score above 0.45 was considered significant (Berger et al., 2008).

ChIP/ChIP-Seq for BNC2

ChIP was performed as previously described (Gomes et al., 2006) using a validated BNC2 antibody (Sigma Atlas) (Figure 14B). In brief, iOSE11 or iFTSEC283 cells grown to 70% confluence in media with a 1:1 ratio of MCDB105/Medium 199, 15% Fetal Bovine Serum (FBS), 10 ng/mL Epidermal Growth Factor, 0.5 µg/mL hydrocortisone, 5 µg/mL insulin, and 34 µg/mL Bovine Pituitary Extract were cross-linked with 1% Formaldehyde in PBS. Crosslinking is quenched by adding Glycine to a concentration of .125M. After washing, cells are collected in Szaks' RIPA buffer (150 mM NaCl, 1% NP-40, 0.5% deoxycholate, 0.1% SDS, 50 mM Tris HCl pH8, 5 mM EDTA, Protease Inhibitors, 50 mM NaF, 0.2 mM sodium orthovanadate, 0.5 mM PMSF) and the lysate is brought to approximately 1 mg/mL. The lysate is then sonicated in Biogenode Sonicating Water Bath for 12 cycles of 30 seconds on and 30 seconds off for 8 minutes. One mg of protein is then mixed with 40 µL of 50% slurry protein A/G agarose beads (Santa Cruz) previously washed in Szaks' RIPA buffer and pre-cleared for 1-2 hours at 4°C. Pre-cleared lysate is then mixed with 5 µg of BNC2 antibody (Sigma Atlas) and 40 µL of 50% slurry protein A/G agarose beads previously washed in Szaks' RIPA buffer and saturated with 1 mg/mL BSA. The mix was incubated overnight at 4°C while rotating. Beads are then washed twice with Szaks' RIPA Buffer, four times with Szaks' IP wash buffer (100 mM Tris HCl pH 8.5, 500 mM LiCl, 1% NP-40, 1% deoxycholate), twice again with Szak' RIPA Buffer and twice with cold TE. Immunocomplexes are eluted by incubating samples at 65°C for 10 minutes in 1.5X Talianidis Elution Buffer (70 mM Tris HCl pH 8, 1 mM EDTA, 1.5% SDS). Crosslinks were reversed by bringing samples to 200 mM NaCl solution and incubating at 65°C for 5 hours. DNA was purified by phenol-chloroform extraction and re-suspended in 50 µL 10 mM Tris pH 8.0.

Real time qPCR was performed using Sybr Green chemistry with primers at the -2184, -914, and -582 positions relative to the TSS (Table 5) in an Applied Biosystems 7900HT Fast Real Time PCR System using absolute quantification with genomic DNA as a standard control to measure the amount of DNA bound by BNC2 in comparison to the IgG control. Data were normalized by taking the percentage of DNA for each site of the highest bound site. ChIP for each cell line was performed in four biological replicates.

For BNC2 ChIP-Seq, four individual ChIP samples were pooled for each cell line (iOSE11 and iFTSEC283) in two biological replicates. Immunoprecipitated DNA was used to generate a sequencing library using the NuGEN Ovation Ultralow Library System with indexed adapters (NuGEN, Inc., San Carlos, CA). The library was PCR amplified and size-selected using AxyPrep Fragment Select beads (Corning Life

Sciences – Axygen Inc., Union City, CA). The size and quality of the library was evaluated using the Agilent BioAnalzyer, and the library was quantitated with the Kapa Library Quantification Kit (Kapa Biosystems, Woburn MA). Each enriched DNA library was then sequenced on an Illumina HiScan SQ sequencer to generate 20-30 million 100-bp-end reads. The raw sequence data was de-multiplexed using the Illumina CASAVA 1.8.2 software (Illumina, Inc., San Diego, CA) and binding sites were identified using the MACS software (Zhang et al., 2008) using input DNA as a control, and a band with setting of 300. All other parameters were set to defaults. The .bam and .wig files were visualized and inspected using the UCSC genome browser (Kent et al., 2002). Peaks used for further analysis had an intensity greater than 0.05 (reads/length), number of reads greater than 50, and a fold change compared to the input greater than 10.

To identify target genes, bed files were uploaded into Galaxy Cistrome (Liu et al., 2011) and the peak2gene Peak Center Annotation tool was used on both the iOSE11 and iFTSEC283 BNC2 ChIP-Seq files to generate a list of genes within 30 kb of the peak centers. BNC2 ChIP-seq peaks for each of the iOSE11 and iFTE283 samples were ranked by their MACS *p*-values, and the top 2000 peaks with the most significant *p*-values were selected for *de novo* motif discovery. The 500 bp sequences surrounding the summits of the top 2000 peaks were extracted, and *de novo* motif discovery was performed for iOSE11 and iFTE283 samples separately using MEME-ChIP (Machanick and Bailey, 2011) with the following parameters: ZOOPS mode, minimum MEME width 6, maximum MEME width 30, maximum 5 MEME motifs, and DREME E-value cutoff 0.05. The top scoring MEME-ChIP motif for both samples were almost identical, and thus were averaged. Enrichment of this motif at the peak summits was examined using CentriMo (Bailey and Machanick, 2012) for iOSE11 and iFTE283 samples separately, as central enrichment is often associated with direct binding of the protein to DNA (Bailey and Machanick, 2012).

DNA-binding preferences were predicted for each of the three ZF pairs of BNC2 (ZF1-2, ZF3-4, and ZF5-6) using B1H-RC (Najafabadi et al., 2015), which consists of a set of Random Forest models that take the protein sequence as input and predict nucleotide preferences at different DNA positions. These predictions were compared to the PBM motifs obtained for the same ZF pairs, as well as to the *de novo* motif obtained from full-length BNC2 ChIP-seq.

Nanostring

pNTAP-BNC2 (or the empty vector) was transfected with Fugene 6 into 293FT cells grown to 70% confluence in DMEM and 10% FBS. Cells were harvested after 24 hours and *BNC2* overexpression was confirmed by Western blotting (Figure 17). RNA was isolated using Trizol RNA Isolation (Life Technoligies) and cleaned using Qiagen RNeasy Mini Kit (Qiagen). The three biological replicates for 293FT cells with the empty vector or over-expressed *BNC2* were applied to a Nanostring platform containing probes for 86 genes and controls (Table 3). A custom NanoString nCounter Gene Expression (GX) CodeSet with probes representing 97 genes was developed and the sample was processed on the NanoString nCounter Analysis System according to the manufacturer's protocol (NanoString Technologies, Seattle WA). The resulting .RCC files containing raw counts were checked for quality in the NanoString nSolver Analysis

Software v1.1, and then exported for normalization and analysis. TTI2, PRDM4, STK35, TOX4, INO80, GIGYF2, and SMG5 were used as reference genes to normalize the data in the NanoString nSolver Analysis Software v 1.1. These genes had a %CV < 50. Normalized data is then analyzed using Graph Pad Prism 6. Genes were considered to be differentially expressed had a p-value <.05 (two tailed t-test).

Tandem Affinity Purification coupled to LC-MS/MS

Full length *BNC2* cDNA was amplified using primers containing *Bam*HI and *Sal*I restriction enzyme sites (Table 5) and cloned into pNTAP-B vector cut with *Bam*HI and *Xho*I. Construct 1-236 was obtained by cutting the 1-524 PCR amplification product with *Bam*HI and *Xho*I (site at 710 bp). Log growing 293FT (1 x 10^8 cells) grown in DMEM plus 10% FBS were transfected with 200 µg pNTAP-BNC2 construct using the calcium phosphate method, as previously described (Swift et al., 2001). Purification of TAP-tagged BNC2 complexes was performed using the InterPlay TAP purification kit (Stratagene) (Woods et al., 2012). A TAP-GFP construct was used as a control to reduce false positive interactions using this purification method.

Following in-gel tryptic digestion, peptides were extracted and concentrated under vacuum centrifugation. A nanoflow liquid chromatograph (Easy-nLC, Proxeon, Odense, Denmark) coupled to an electrospray ion trap mass spectrometer (LTQ, Thermo, San Jose, CA) was used for tandem mass spectrometry peptide sequencing experiments. The sample was first loaded onto a trap column (BioSphere C18 reversed-phase resin, 5µm, 120Å, 100 µm ID, NanoSeparations, Nieuwkoop, Netherlands) and washed for 3 minutes at 8 mL / minute. The trapped peptides were eluted onto the analytical column, (BioSphere C18 reversed-phase resin, 150mm, 5µm, 120Å, 100 µm ID, NanoSeparations, Nieuwkoop, Netherlands). Peptides were eluted in a 30 minute gradient from 5% B to 45% B (solvent A: 2% acetonitrile + 0.1% formic acid; solvent B: 90% acetonitrile + 0.1% formic acid) with a flow rate of 300 nL/minute. Five tandem mass spectra were collected in a data-dependent manner following each survey scan. Sequences were assigned using Mascot (www.matrixscience.com) searches against human Swiss Prot entries (Sprot_20090505, 20402 entries). Carbamidomethylation of cysteine, methionine oxidation, and deamidation of asparagine and glutamine were selected as variable modifications, and as many as 2 missed tryptic cleavages were allowed. Precursor mass tolerance is set to 2.5 and fragment ion tolerance to 0.8. Results from Mascot were compiled in Scaffold, which was used for manual inspection of peptide assignments and protein identifications.

Database searches were conducted against human entries in the SwissProt database (v.20090505) using Mascot (Matrix Science, London, UK; version 2.2.04) (Perkins et al., 1999), assuming the digestion enzyme trypsin and allowing as many as 2 missed cleavages. Tandem mass spectra were matched to peptide sequences with a peptide ion mass tolerance of 1.2 Da and fragment ion mass tolerance of 0.80 Da. Oxidation of methionine and carbamidomethylation of cysteine were specified as variable modifications. Assignments were manually verified by inspection of the tandem mass spectra and coalesced into Scaffold reports (v.2.0, available at www.proteomesoftware.com) for statistical analysis and data presentation.

Scaffold (version Scaffold_2_04_00, Proteome Software Inc., Portland, OR) was used to validate MS-MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability as specified by the Peptide Prophet algorithm (Keller et al., 2002). Protein identifications were accepted if they could be established at greater than 50.0% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii et al., 2003). Proteins that contained similar peptides and could not be differentiated based on MS-MS analysis alone were grouped to satisfy the principles of parsimony. Proteins appearing in control GFP purification and in the BNC2 samples were removed from the final datasets. Final protein list underwent gene ontology enrichment analysis from the WebGestalt program using h_sapiens genome as a reference set (Zhang et al., 2005).

Transcriptional Repression Assay

Fragments as well as full length *BNC2* were PCR amplified from plasmid containing *BNC2* cDNA using primers to clone in frame to GAL4 DBD (Table 5). These plasmids were then co-transfected with pGAL4-TK-Luc (Yang et al., 2001) expressing firefly luciferase and pRL-SV40 expressing *Renilla* luciferase as an internal control into 293FT cells. Luciferase levels were measured 24 hours post-transfection using Dual Luciferase II Assay Kit (Promega).

Table 5: List of oligos used in this study:

Primer Name	Sequence	Assay
T5 FWD	attb-TAAGTAGAGACGGGGTTTCA	Tiling Clones
T5 REV	attb-CTGATGGACCATTCTTCACT	Tiling Clones
T6 FWD	attb-GGTGGAAAGCAAACTAAATG	Tiling Clones
T6 REV	attb-TAGTTCTGTTGTGCAGGTTG	Tiling Clones
T7.1 FWD	attb-TGGGGGTTTTCATTGCCAGG	Tiling Clones

Table 5 (Continued)

Primer Name	Sequence	Assay
T7.1 REV	attb-GACTCGACGATGTGCTGTCC	Tiling Clones
T7.2 FWD	attb-GTGAAGCTGCACAGACACTA	Tiling Clones
T7.2 REV	attb-CAGAATGTTGCACAAAAAGA	Tiling Clones
T7.3 FWD	attb-TCTTTTTGTGCAACATTCTG	Tiling Clones
T7.3 REV	attb-ATTAAGTTGGGTGGTGTTTG	Tiling Clones
T7.4 FWD	attb-ACCGTGCTGAACCCTTAACT	Tiling Clones
T7.4 REV	attb-GAGCCCAGGACTGTGGTTAC	Tiling Clones
T7.5 FWD	attb-CTCTGCTTTTGTCTGCTTCT	Tiling Clones
T7.5 REV	attb-GGACCTACGGGAACTTTTAC	Tiling Clones
T7.6 FWD	attb-ATTCCGAATGTGAAGACAAG	Tiling Clones
T7.6 REV	attb-TTTTCACTAGGAACCGGTAA	Tiling Clones
T7.7 FWD	attb-GTGCAAGCCCCACAAGTTTT	Tiling Clones
T7.7 REV	attb-GATGCAACCTGTCCCCAGAA	Tiling Clones
T7.8 FWD	attb-TCTCCGAGTTATGCAGATTT	Tiling Clones
T7.8 REV	attb-GAGCTTTGCAAGTTAGAGGA	Tiling Clones
T8 FWD	attb-AGAGACAACCCAAGATAGCA	Tiling Clones
T8 REV	attb-CCAATGACGAAATGTATGTG	Tiling Clones
T11 FWD	attb-GGGTGGGGGTGAGGATGATA	Tiling Clones
T11 REV	attb-TTTGCCTGTAGTGGGTGCTC	Tiling Clones
T12 FWD	attb-TGCCTGGCTTAGTCTTTATT	Tiling Clones
T12 REV	attb-AGGAGAAGGAATAGCTGCTT	Tiling Clones
TC FWD	attb-GAATATGACTGGCACCACTT	Tiling Clones
TC REV	attb-AATAAAGACTAAGCCAGGCA	Tiling Clones
rs62541878 sense	tatctctacaaaatatatatatatatatatataaatttaccaggcatcgtggcttgc	Site Directed Mutagenesis
rs62541878 antisense	gcaagccacgatgcctggtaaatttatatatatatatatatatattttgtagagata	Site Directed Mutagenesis
rs62541920 sense	atacatacacagtgagtcatttaagagtttcacattctgccttc	Site Directed Mutagenesis
rs62541920 antisense	gaaggcagaatgtgaaactcttaaatgactcactgtgtatgtat	Site Directed Mutagenesis
rs12379183 sense	tcaaagagaaaatagagcaaaaagaacaaaactgatgttgttatgtacggatattt	Site Directed Mutagenesis
rs12379183 antisense	aaatatccgtacataacaacatcagttttgttctttttgctctattttctctttga	Site Directed Mutagenesis
rs10962647 sense	tcagctctgcttttgtctgctgctttttgtaatcacatatctc	Site Directed Mutagenesis
rs10962647 antisense	gagatatgtgattacaaaaagcagcagacaaaagcagagctga	Site Directed Mutagenesis
rs62541922 sense	ccagccgccggcccctcactcgg	Site Directed Mutagenesis
rs62541922 antisense	ccgagtgaggggccggcggctgg	Site Directed Mutagenesis
rs62541923 sense	ggtccccggccagccctcctcag	Site Directed Mutagenesis
rs62541923 antisense	ctgaggagggctggccggggacc	Site Directed Mutagenesis
rs200648906 sense	cagctgtcacacacacacgaaaaaaaaattgcggggc	Site Directed Mutagenesis
rs200648906 antisense	gccccgcaatttttttttcgtgtgtgtgtgacagctg	Site Directed Mutagenesis

Table 5 (Continued)

Primer Name	Sequence	Assay
rs10810657 sense	cttcgttacacagcatatatctgcacccctgcc	Site Directed Mutagenesis
rs10810657 antisense	ggcaggggtgcagatatatgctgtgtaacgaag	Site Directed Mutagenesis
rs12350739 sense	Gtgctgctgatctcatgcccttcctctgg	Site Directed Mutagenesis
rs12350739 antisense	Ccagaggaagggcatgagatcagcagcac	Site Directed Mutagenesis
rs77507622 sense	Attccagaaatcattattaggcagtttcttagagcaattcatgggtt	Site Directed Mutagenesis
rs77507622 antisense	aacccatgaattgctctaagaaactgcctaataatgatttctggaat	Site Directed Mutagenesis
rs10810657_Effect A	ATCACTTCGTTACACAGCATATATCTGCACCCCTGCCGGCA	EMSA
rs10810657_Effect T	TGCCGGCAGGGGTGCAGATATATGCTGTGTAACGAAGTGAT	EMSA
rs10810657_Reference T	ATCACTTCGTTACACAGCATTTATCTGCACCCCTGCCGGCA	EMSA
rs10810657_Reference A	TGCCGGCAGGGGTGCAGATAAATGCTGTGTAACGAAGTGAT	EMSA
rs77507622_Reference A	CCAGAAATCATTATTAGGCAATTTCTTAGAGCAATTCATGG	EMSA
rs77507622_Reference T	CCATGAATTGCTCTAAGAAATTGCCTAATAATGATTTCTGG	EMSA
rs77507622_Effect G	CCAGAAATCATTATTAGGCAGTTTCTTAGAGCAATTCATGG	EMSA
rs77507622_Effect C	CCATGAATTGCTCTAAGAAACTGCCTAATAATGATTTCTGG	EMSA
rs12350739_Effect A	TCATCAGTGCTGCTGATCTCATGCCCTTCCTCTGGCAAACC	EMSA
rs12350739_Effect T	GGTTTGCCAGAGGAAGGGCATGAGATCAGCAGCACTGATGA	EMSA
rs12350739_Reference G	TCATCAGTGCTGCTGATCTCGTGCCCTTCCTCTGGCAAACC	EMSA
rs12350739_Reference C	GGTTTGCCAGAGGAAGGGCACGAGATCAGCAGCACTGATGA	EMSA
rs113780397_Effect A	AGATTGAGCCACTGCACTGCATCCTGGGTGACAGAGCGAGA	EMSA
rs113780397_Effect T	TCTCGCTCTGTCACCCAGGATGCAGTGCAGTGGCTCAATCT	EMSA
rs113780397_Reference G	AGATTGAGCCACTGCACTGCGTCCTGGGTGACAGAGCGAGA	EMSA
rs113780397_Reference C	TCTCGCTCTGTCACCCAGGACGCAGTGCAGTGGCTCAATCT	EMSA
rs117224476_Reference T	TATATTATATATAATGTATATATTATATATTATATAATATA	EMSA
rs117224476_Reference A	TATATTATATAATATATAATATATACATTATATATAATATA	EMSA
rs117224476_Effect G	TATATTATATATAATGTATAGATTATATATTATATAATATA	EMSA
rs117224476_Effect C	TATATTATATAATATATAATCTATACATTATATATAATATA	EMSA
rs12379183_Reference A	GCAAAAAGAACAAAACTGATATTGTTATGTACGGATATTTT	EMSA
rs12379183_Reference T	AAAATATCCGTACATAACAATATCAGTTTTGTTCTTTTTGC	EMSA
rs12379183_Effect G	GCAAAAAGAACAAAACTGATGTTGTTATGTACGGATATTTT	EMSA
rs12379183_Effect C	AAAATATCCGTACATAACAACATCAGTTTTGTTCTTTTTGC	EMSA
rs62541920_Reference G	CATACACAGTGAGTCATTTAGGAGTTTCACATTCTGCCTTC	EMSA
rs62541920_Reference C	GAAGGCAGAATGTGAAACTCCTAAATGACTCACTGTGTATG	EMSA
rs62541920_Effect A	CATACACAGTGAGTCATTTAAGAGTTTCACATTCTGCCTTC	EMSA
rs62541920_Effect T	GAAGGCAGAATGTGAAACTCTTAAATGACTCACTGTGTATG	EMSA
rs10962647_Reference T	CAGCTCTGCTTTTGTCTGCTTCTTTTTGTAATCACATATCT	EMSA
rs10962647_Reference A	AGATATGTGATTACAAAAAGAAGCAGACAAAAGCAGAGCTG	EMSA
rs10962647_Effect G	CAGCTCTGCTTTTGTCTGCTGCTTTTTGTAATCACATATCT	EMSA
rs10962647_Effect C	AGATATGTGATTACAAAAAGCAGCAGACAAAAGCAGAGCTG	EMSA
rs3814113_Major T	CAGGGTACCTGCTCCATATCTTCTGGACCAGTTCTCCAAAC	EMSA

Table 5 (Continued)

Primer Name	Sequence	Assay
rs3814113_Major A	GTTTGGAGAACTGGTCCAGAAGATATGGAGCAGGTACCCTG	EMSA
rs3814113_minor C	CAGGGTACCTGCTCCATATCCTCTGGACCAGTTCTCCAAAC	EMSA
rs3814113_minor G	GTTTGGAGAACTGGTCCAGAGGATATGGAGCAGGTACCCTG	EMSA
rs181552334_Major A	ATATATATAATATATATTACATAATATATAATATATATAAT	EMSA
rs181552334_Major T	ATTATATATATTATATATTATGTAATATATATTATATATAT	EMSA
rs181552334_minor G	ATATATATAATATATATTACGTAATATATAATATATATAAT	EMSA
rs181552334_minor C	ATTATATATATTATATATTACGTAATATATATTATATATAT	EMSA
BPR1F	CGCTACCACCACCACATACAT	3C
BPAR1F	GCTCTGAACACGCACAGACA	3C
BPR2R	GGTGGTGCACACCTGTAGAG	3C
BPAR2R	GGGCAATGAGCTGTGTCTCT	3C
9p22BNCEcoF1	ACTTTTGGGTAAAGAGGGACAA	3C
9p22BNCpEcoF2	GCTGGCTTGATGCTATTCCC	3C
9p22BNCpEcoF3	AGCAATTTGTGAAGTACCAGGC	3C
9p22BNCEcoR4	GGTATAGTGAGAAGGGCACCA	3C
9p22BNCEcoR5	AACTGAGACCAGACCACAAGT	3C
9p22BNCEcoF6	CACCGCACCGATCCTGTTT	3C
9p22BNCEcoF7	ACTCCAGTCTGGGCAACAAG	3C
9p22BNCEcoF8	GTCATGAGATCGTGGCTGGG	3C
9p22BNCEcoF9	GGCGTGAAACTTCTTGTTATGTGA	3C
9p22BNCEcoF10	GACTGAGCCTGAAGGAAGGC	3C
9p22BNCEcoF11	ACAGATACCAAGTGCAAACTGC	3C
9p22BNCEcoF12	GGGAGGCTGAGACATGAGAC	3C
9p22BNCEcoR13	CCCCATCTGGGACTTGAAGG	3C
9p22BNCEcoR14	CACCAATAAGCGATCAGCTCC	3C
9p22BNCEcoF15	CAGGGCCAAGAATCTACCGC	3C
9p22BNCEcoR16	GAAAAATCACCTGTGTGGGCA	3C
9p22BNCEcoF17	GCACAAGGCCCTTATTCCCA	3C
9p22BNCEcoF18	TGCCTCTGCCAGAATGATGT	3C
9p22BNCEcoR19	TGTGTCATTGAGTGGTGTTGAT	3C
9p22BNCEcoR20	ATCTCTTGAAGCCAGCCATTT	3C
9p22BNCEcoR21	TGTGAGAGTGCCTCGGTGTA	3C
9p22BNCEcoF22	GGCATGCTGCCACATATTCAG	3C
9p22BNCEcoF23	ACCACAGAGAAGGTGGCAAG	3C
9p22BNCEcoR24	TGTTTCCCTCTCCTCCCCAA	3C
9p22CNTLNEcoF1	CATAGGAGATATACATCAGTTGCCA	3C
9p22CNTLNEcoF2	TGGTGCTTGTAGAGGGGTTTC	3C
9p22CNTLNEcoF3	TGGAGACAGGGTAGCGATCA	3C
9p22CNTLNEcoF4	GCTTAGCACTGGACTCAGCA	3C
9p22CNTLNEcoF5	ACCCAACAAGGCTTGAAACA	3C

Table 5 (Continued)

Primer Name	Sequence	Assay
9p22CNTLNEcoF6	GGGTTGCTAGAGACTTGGGG	3C
9p22CNTLNEcoF7	ATCTCCTCAGTGGCCTTTGT	3C
9p22CNTLNEcoR8	GCCGAGCCCTCATGATGTAA	3C
1-236 FWD	TA*GGATCC*AGATGGCACACCTTGGGCCCAC	GAL4-BNC2 Constructs
1-524 FWD	TA*GGATCC*AGATGGCACACCTTGGGCCCAC	GAL4-BNC2 Constructs
1-524 REV	CT*GTCGAC*ACTTGCTATGACAGGGGTGG	GAL4-BNC2 Constructs
385-524 FWD	CA*GGATCC*GAAATGCCCTGACCAGCATTAC	GAL4-BNC2 Constructs
385-524 REV	CT*GTCGAC*ACTTGCTATGACAGGGGTGG	GAL4-BNC2 Constructs
519-818 FWD	CA*GGATCC*CCACCCCTGTCATAGCAAGT	GAL4-BNC2 Constructs
519-818 REV	CT*GTCGAC*TTTGAGGGCTGCCATAATTC	GAL4-BNC2 Constructs
818-1099 FWD	GA*GGATCC*TGAATTATGGCAGCCCTCAAA	GAL4-BNC2 Constructs
818-1099 REV	AA*GTCGAC*CTAATCTACTGAAGTGAAGG	GAL4-BNC2 Constructs
951-1099 FWD	GA*GGATCC*GGAGAGGCATGGCAGAGGACTA	GAL4-BNC2 Constructs
951-1099 REV	AA*GTCGAC*CTAATCTACTGAAGTGAAGG	GAL4-BNC2 Constructs
1-1099 FWD	TA*GGATCC*AGATGGCACACCTTGGGCCCAC	GAL4-BNC2 Constructs
1-1099 REV	AA*GTCGAC*CTAATCTACTGAAGTGAAGG	GAL4-BNC2 Constructs
BNC2 ZF 1,2 FWD	attb-CAGAATTCTGCCCCAGTCAGTG	ZF Protein Constructs
BNC2 ZF 1,2 REV	attb-TACTAGCTGGCTAGGGAGAGGA	ZF Protein Constructs
BNC2 ZF 3,4 FWD	attb-GACATGTTTTACATGAGCCAGT	ZF Protein Constructs
BNC2 ZF 3,4 REV	attb-GTTCAGGTGGGAGTCTTCAGTC	ZF Protein Constructs
BNC2 ZF 5,6 FWD	attb-AGAGGCATGGCAGAGGACTA	ZF Protein Constructs
BNC2 ZF 5,6 REV	attb-ATCTACTGAAGTGAAGGGAA	ZF Protein Constructs
-2184 FWD	CCTGCAGATGCAACCTGTCCCC	Site Specific ChIP
-2184 REV	TCTGCATTCGTGGATTCTGTGCAT	Site Specific ChIP
-914 FWD	GCACAAAACGCTCCGCCACC	Site Specific ChIP
-914 REV	GGCGGAGGAAAACCCAGCGG	Site Specific ChIP
-582 FWD	TTCCTCGGCGTTTCGCAGCC	Site Specific ChIP
-582 REV	GCGGGCGTGGAGGTAGAGGT	Site Specific ChIP

CHAPTER FIVE:

DISCUSSION

Note to Reader

A manuscript that has been submitted for review includes portions of this chapter.

Here, we start from EOC GWAS findings and delineate a potential mechanistic basis for susceptibility at the 9p22.2 locus (Figure 19). We identify allele specific effects for 5 candidate functional SNPs in three genomic regions, 7, 8 and 11, which contribute to the regulation of *BNC2,* a pleiotropic gene encoding a TF involved in ovarian development, skin pigmentation, and height in fish, mice, and humans.

In Chapter two we identified 5 functional SNPs with allele specific function by assessing their regulatory potential in several different assays. These assays included FAIRE-seq and ChIP-seq for histone markers to identify regulatory regions in the locus as well as luciferase reporter assays measuring said region’s transcription activity. Allele specific functions were measured with luciferase reporter assays comparing transcription activity between plasmids with the major and minor allele. EMSA discern which SNPs have allele specific nuclear extract binding. These assays measure different yet important mechanisms of transcription therefore we have confidence that these 5 positive SNPs truly represent the causal SNPs in the 9p22.2 locus.

Functional analysis of disease susceptibility loci has mainly focused on associated SNPs that locate to enhancers due to the prevalence of enhancers in the genome. Yet, evidence for a functional role as an enhancer for Region 11 is weak; it is not conserved, is highly repetitive, and has few enhancer activity-associated chromatin features (Figure 10, Figure 8A). Consistent with the data, it displayed significant luciferase activity in only one experiment (Figure 8B). It failed to show significant 3C interactions to *BNC2* or *CNTLN* promoters in chapter three (Figure 12C). Moreover, inspection of histone modifications by ChIP-seq from ENCODE for over 70 cell lines failed to reveal any chromatin markers associated with enhancers. Thus, it is unlikely that this region acts as an enhancer even in a different cell type.

In silico analysis suggests that the region may harbor a scaffold/matrix attachment region (S/MAR). S/MARs have been proposed to assist in chromatin looping, maintain the local 3D structure of the genome and modulate gene expression (Linnemann et al., 2009). Importantly, polymorphisms in S/MARs can affect their ability to attach to the nuclear scaffold/matrix (Kisseljova et al., 2014). It is possible that S/MARs, which are not well characterized, may play a significant role in the association with ovarian cancer risk. Quantitative PCR for region 11 in DNA purified from a nuclear scaffold extract could confirm whether region 11 is indeed a S/MAR. Knocking out region 11 and looking at changes in expression of potential downstream targets would identify the target gene of region 11. As mentioned earlier, S/MARs are not well characterized therefore changes in DNA binding to the nuclear scaffold may have an entirely unique effect on development and cancer oncogenesis outside of the realm of transcription regulation.

In Chapter Three, we identified *BNC2* as the major target of the 9p22 causal SNPs. In ovarian cells, the promoter of *BNC2* is active and transcripts for *BNC2* are present. Additionally two of the causal SNPs in Region 7 implicate *BNC2* as a target gene as they are located in its major promoter. Two of the causal SNPs in Region 8 overlap with enhancer marks and physically interact with the *BNC2* promoter. Thus, these regions, conserved in mouse, contribute to the regulation of *BNC2* expression (Figure 10).

Our data indicate that the mechanism by which genetic variation at the locus affect ovarian cancer susceptibility may be mediated primarily by multiple non-coding elements including enhancer and promoter elements which target *BNC2* and by a putative S/MAR element. Limitations of this work include the possibility that enhancer landscapes may change significantly during development (Pennacchio et al., 2006), the limited regulatory profiling information that has been performed in ovarian cells (*e.g.* lack of CTCF repressor marks and putative non-coding RNA elements data) and the possibility of false positive and false negative findings.

However, several lines of evidence suggest that we have captured the functional features relevant to ovarian cancer risk. First, our use of a large panel of ovarian normal and cancer cell lines derived from different origins provide a broad view of regulatory activity. Second, most (~90%) enhancers are linked to single promoter in primary mouse cells (Kieffer-Kwon et al., 2013) supporting the hypothesis that Region 8 targets *BNC2* exclusively although it is unclear how other regulatory features present in the locus might interact with the regulatory network described here. For example, large active enhancer clusters that drive expression of cell identity genes, also called super enhancers (Hnisz et al., 2013) are nearby. This paper discusses that the repertoire of cell identity genes mostly includes TFs which in turn use an auto-regulatory loop to regulate their own expression by binding to these super enhancers. This super enhancer (Figure 19B) contains BNC2 binding sites defined by our ChIP-seq data suggesting that *BNC2* represents a cell identity gene or master regulator in ovarian cells.

In Chapter Four we show that *BNC2*, an ancient gene with potentially very important functions in living organisms, acts as a transcription regulator in enhancer regions potentially through recruitment of the NuRD complex of chromatin remodeling (Gunther et al., 2013; Hu and Wade, 2012; Ramirez and Hagman, 2009). We also show that BNC2 targets enhancer regions likely operating to modulate the expression of downstream genes involved in cell-cell and cell-extracellular matrix communication.

Genome-wide and candidate gene association studies have identified significant associations between the 9p22.2 locus and skin pigmentation in Europeans (rs10756819) (Jacobs et al., 2013) and Asians (Hider et al., 2013), and freckling (rs2153271) (Eriksson et al., 2010) (Figure 19B). Importantly, functional dissection identified rs12350739 as the likely causal variant associated with saturation of skin color (Visser et al., 2014). This SNP is one of the functional SNPs identified in the present study mapping to a *BNC2* enhancer (Region 8). The locus also contains a region of Neanderthal DNA (Chr9: 16,720,121-16,786,930) thought to confer adaptive advantage to colder climates by modulation of skin pigmentation (Sankararaman et al., 2014; Vernot and Akey, 2014) (Figure 19B). In addition, the locus has been found to be associated with human height (rs2149163 and rs3927536) (Wood et al., 2014) (Figure

19B). In relation to EOC, the locus is associated to abnormal ovarian ultrasound results (rs12379183, a functional SNP in the present study located in Region 7) (Figure 19B). The link between *BNC2* and ovarian biology is further supported by the results of deletion of *BNC2* in mice and zebrafish. In the *bonaparte* mutants, in addition to the defect in skin pigmentation (lack of body stripes) and stunted growth, the ovaries are dysmorphic leading to infertility (Lang et al., 2009). In mice, *Bnc2* is expressed in ovarian somatic cells such as theca cells and *Bnc2*$^{-/-}$ female mice present with an excess of stromal cells and a greatly reduced number of oocytes leading to infertility (Philippe Djian, personal communication). These mice are also born abnormally small with cleft palates suggesting that *bnc2* plays a role in skeletal development (Vanhoutteghem et al., 2009). This data suggests that the *bnc2* mutants/knock-outs in model organisms influence the same traits identified to be associated with the locus further supporting that *BNC2* is the target gene.

Decreased parity is a major risk factor for ovarian cancer and disruption of *BNC2* in zebrafish and mice indicate its role in fertility. Additionally, evidence suggesting that *BNC2* is an ovarian cell identity gene/master regulator indicates its role in ovarian development and maintenance of function. Therefore *BNC2* may indirectly influence ovarian cancer risk by primarily influencing ovarian development and fertility. Additional functional assays studying ovarian development and the menstrual cycle, such as conditional knock-outs of *BNC2*, may inform the role of *BNC2* in ovarian development and fertility.

In summary, we propose that the mechanism of ovarian cancer susceptibility in the 9p22.2 locus is likely mediated by changes in a transcriptional regulatory network involving several enhancer and promoter elements acting on *BNC2* and a putative S/MAR region*.* While our data strongly suggests that the association signal is mediated through *BNC2* regulation the individual contribution of each element and gene(s) to risk must wait further analysis.

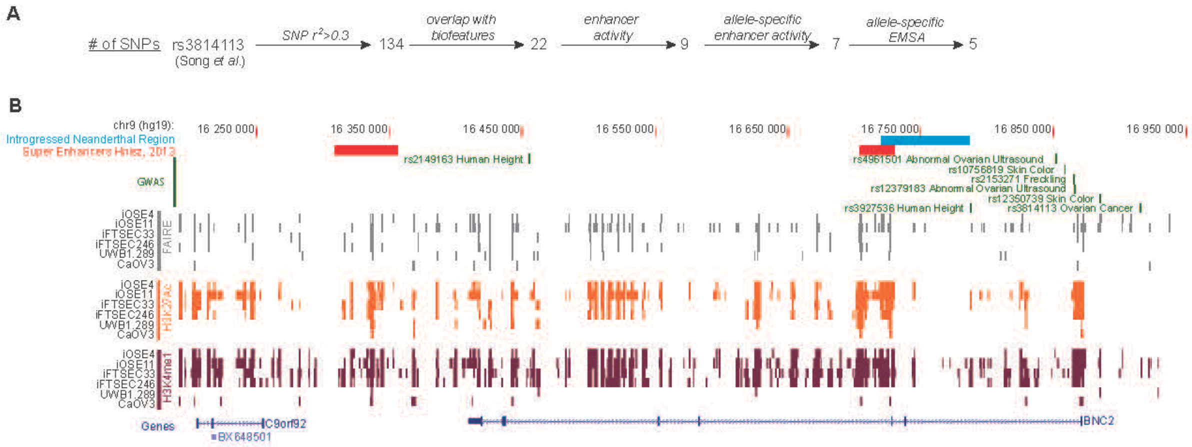


Figure 19: Outline of the study. A. Summary of identifying the functional SNPs. **B.** Further evidence tells us that the 9p22 locus and *BNC2* are important. An introgressed Neanderthal region (light blue bar) lies within the *BNC2* gene and may influence pigmentation. Two super enhancers (red bars) lie on and near the *BNC2* gene and are believed to mainly target TFs involved in cell type identity. Other SNPs at the locus are implicated in abnormal ovarian ultrasound as well as skin color and pigmentation (green).

REFERENCES

Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Ames, B.N., Gold, L.S., and Willett, W.C. (1995). The causes and prevention of cancer. Proceedings of the National Academy of Sciences of the United States of America *92*, 5258-5265.

Antoniou, A.C., Pharoah, P.D., McMullan, G., Day, N.E., Stratton, M.R., Peto, J., Ponder, B.J., and Easton, D.F. (2002). A comprehensive model for familial breast cancer incorporating BRCA1, BRCA2 and other genes. British journal of cancer *86*, 76-83.

Bailey, T.L., and Machanick, P. (2012). Inferring direct DNA binding from ChIP-seq. Nucleic Acids Res *40*, e128.

Barrett, J.C., Fry, B., Maller, J., and Daly, M.J. (2005). Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics *21*, 263-265.

Baysal, B.E., DeLoia, J.A., Willett-Brozick, J.E., Goodman, M.T., Brady, M.F., Modugno, F., Lynch, H.T., Conley, Y.P., Watson, P., and Gallion, H.H. (2004). Analysis of CHEK2 gene for ovarian cancer susceptibility. Gynecologic oncology *95*, 62-69.

Berger, M.F., Badis, G., Gehrke, A.R., Talukder, S., Philippakis, A.A., Pena-Castillo, L., Alleyne, T.M., Mnaimneh, S., Botvinnik, O.B., Chan, E.T.*, et al.* (2008). Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell *133*, 1266-1276.

Berger, M.F., and Bulyk, M.L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc *4*, 393-411.

Berger, M.F., Philippakis, A.A., Qureshi, A.M., He, F.S., Estep, P.W., 3rd, and Bulyk, M.L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol *24*, 1429-1435.

Berns, E.M., and Bowtell, D.D. (2012). The changing view of high-grade serous ovarian cancer. Cancer research *72*, 2701-2704.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R.*, et al.* (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169-181.

Blackwood, E.M., and Kadonaga, J.T. (1998). Going the distance: a current view of enhancer action. Science *281*, 60-63.

Bojesen, S.E., Pooley, K.A., Johnatty, S.E., Beesley, J., Michailidou, K., Tyrer, J.P., Edwards, S.L., Pickett, H.A., Shen, H.C., Smart, C.E.*, et al.* (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nature genetics *45*, 371-384, 384e371-372.

Bolton, K.L., Tyrer, J., Song, H., Ramus, S.J., Notaridou, M., Jones, C., Sher, T., Gentry-Maharaj, A., Wozniak, E., Tsai, Y.Y.*, et al.* (2010). Common variants at 19p13 are associated with susceptibility to ovarian cancer. Nature genetics *42*, 880-884.

Bondarenko, V.A., Jiang, Y.I., and Studitsky, V.M. (2003). Rationally designed insulator-like elements can block enhancer action in vitro. The EMBO journal *22*, 4728-4737.

Bosetti, C., Negri, E., Trichopoulos, D., Franceschi, S., Beral, V., Tzonou, A., Parazzini, F., Greggi, S., and La Vecchia, C. (2002). Long-term effects of oral contraceptives on ovarian cancer risk. International journal of cancer Journal international du cancer *102*, 262-265.

Braem, M.G., Onland-Moret, N.C., van den Brandt, P.A., Goldbohm, R.A., Peeters, P.H., Kruitwagen, R.F., and Schouten, L.J. (2010). Reproductive and hormonal factors in association with ovarian cancer in the Netherlands cohort study. American journal of epidemiology *172*, 1181-1189.

Braman, J., Papworth, C., and Greener, A. (1996). Site-directed mutagenesis using double-stranded plasmid DNA templates. Methods in molecular biology *57*, 31-44.
Britten, R.J., and Davidson, E.H. (1969). Gene regulation for higher cells: a theory. Science *165*, 349-357.

Brosens, I.A., and Vasquez, G. (1976). Fimbrial microbiopsy. The Journal of reproductive medicine *16*, 171-178.

Bu, S.Z., Yin, D.L., Ren, X.H., Jiang, L.Z., Wu, Z.J., Gao, Q.R., and Pei, G. (1997). Progesterone induces apoptosis and up-regulation of p53 expression in human ovarian carcinoma cell lines. Cancer *79*, 1944-1950.

Bushey, A.M., Dorman, E.R., and Corces, V.G. (2008). Chromatin insulators: regulatory mechanisms and epigenetic inheritance. Molecular cell *32*, 1-9.

Cairns, B.R. (2009). The logic of chromatin architecture and remodelling at promoters. Nature *461*, 193-198.

Cancer Genome Atlas Research, N. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.

Cardon, L.R., and Bell, J.I. (2001). Association study designs for complex diseases. Nature reviews Genetics *2*, 91-99.

Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L., and Nickerson, D.A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. American journal of human genetics *74*, 106-120.

Carlson, J.W., Miron, A., Jarboe, E.A., Parast, M.M., Hirsch, M.S., Lee, Y., Muto, M.G., Kindelberger, D., and Crum, C.P. (2008). Serous tubal intraepithelial carcinoma: its potential role in primary peritoneal serous carcinoma and serous cancer prevention. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *26*, 4160-4165.

Casadei, S., Norquist, B.M., Walsh, T., Stray, S., Mandell, J.B., Lee, M.K., Stamatoyannopoulos, J.A., and King, M.C. (2011). Contribution of inherited mutations in the BRCA2-interacting protein PALB2 to familial breast cancer. Cancer research *71*, 2222-2229.

Castera, L., Krieger, S., Rousselin, A., Legros, A., Baumann, J.J., Bruet, O., Brault, B., Fouillet, R., Goardon, N., Letac, O*., et al.* (2014). Next-generation sequencing for the diagnosis of hereditary breast and ovarian cancer using genomic capture targeting multiple candidate genes. European journal of human genetics : EJHG *22*, 1305-1313.

Chen, K., Ma, H., Li, L., Zang, R., Wang, C., Song, F., Shi, T., Yu, D., Yang, M., Xue, W*., et al.* (2014). Genome-wide association study identifies new susceptibility loci for epithelial ovarian cancer in Han Chinese women. Nature communications *5*, 4682.

Chiaravalli, A.M., Furlan, D., Facco, C., Tibiletti, M.G., Dionigi, A., Casati, B., Albarello, L., Riva, C., and Capella, C. (2001). Immunohistochemical pattern of hMSH2/hMLH1 in familial and sporadic colorectal, gastric, endometrial and ovarian carcinomas with instability in microsatellite sequences. Virchows Archiv : an international journal of pathology *438*, 39-48.

Coetzee, S.G., Shen, H.C., Hazelett, D.J., Lawrenson, K., Kuchenbaecker, K., Tyrer, J., Rhie, S.K., Levanon, K., Karst, A., Drapkin, R*., et al.* (2015). Cell Type Specific Enrichment Of Risk Associated Regulatory Elements At Ovarian Cancer Susceptibility Loci. Human molecular genetics.

Colhoun, H.M., McKeigue, P.M., and Davey Smith, G. (2003). Problems of reporting genetic associations with complex outcomes. Lancet *361*, 865-872.

Corbi, N., Libri, V., Fanciulli, M., and Passananti, C. (1998). Binding properties of the artificial zinc fingers coding gene Sint1. Biochemical and biophysical research communications *253*, 686-692.

Corbi, N., Perez, M., Maione, R., and Passananti, C. (1997). Synthesis of a new zinc finger peptide; comparison of its 'code' deduced and 'CASTing' derived binding sites. FEBS letters *417*, 71-74.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nature genetics *46*, 1311-1320.

Couch, F.J., Wang, X., McGuffog, L., Lee, A., Olswold, C., Kuchenbaecker, K.B., Soucy, P., Fredericksen, Z., Barrowdale, D., Dennis, J.*, et al.* (2013). Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS genetics *9*, e1003212.

Coulet, F., Fajac, A., Colas, C., Eyries, M., Dion-Miniere, A., Rouzier, R., Uzan, S., Lefranc, J.P., Carbonnel, M., Cornelis, F.*, et al.* (2013). Germline RAD51C mutations in ovarian cancer susceptibility. Clinical genetics *83*, 332-336.

Cramer, D.W., Barbieri, R.L., Fraer, A.R., and Harlow, B.L. (2002). Determinants of early follicular phase gonadotrophin and estradiol concentrations in women of late reproductive age. Human reproduction *17*, 221-227.

Dahlman, I., Eaves, I.A., Kosoy, R., Morrison, V.A., Heward, J., Gough, S.C., Allahabadia, A., Franklyn, J.A., Tuomilehto, J., Tuomilehto-Wolf, E.*, et al.* (2002). Parameters for reliable results in genetic association studies in common disease. Nature genetics *30*, 149-150.

Dekker, J. (2006). The three 'C' s of chromosome conformation capture: controls, controls, controls. Nature methods *3*, 17-21.

Dellorusso, C., Welcsh, P.L., Wang, W., Garcia, R.L., King, M.C., and Swisher, E.M. (2007). Functional Characterization of a Novel BRCA1-Null Ovarian Cancer Cell Line in Response to Ionizing Radiation. Molecular cancer research : MCR *5*, 35-45.

Dijkwel, P.A., and Hamlin, J.L. (1999). Physical and genetic mapping of mammalian replication origins. Methods *18*, 418-431.

Domcke, S., Sinha, R., Levine, D.A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. Nature communications *4*, 2126.

Dreher, D., and Junod, A.F. (1996). Role of oxygen free radicals in cancer development. European journal of cancer *32A*, 30-38.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R.*, et al.* (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Elrod-Erickson, M., Rould, M.A., Nekludova, L., and Pabo, C.O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure *4*, 1171-1180.

Emerson, R.O., and Thomas, J.H. (2009). Adaptive evolution in zinc finger transcription factors. PLoS genetics *5*, e1000325.

Eriksson, N., Macpherson, J.M., Tung, J.Y., Hon, L.S., Naughton, B., Saxonov, S., Avey, L., Wojcicki, A., Pe'er, I., and Mountain, J. (2010). Web-based, participant-driven studies yield novel genetic associations for common traits. PLoS genetics *6*, e1000993.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M.*, et al.* (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43-49.

Fathalla, M.F. (1971). Incessant ovulation--a factor in ovarian neoplasia? Lancet *2*, 163.
Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M.J.*, et al.* (2009). Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. The New England journal of medicine *361*, 123-134.

Ford, D., Easton, D.F., Stratton, M., Narod, S., Goldgar, D., Devilee, P., Bishop, D.T., Weber, B., Lenoir, G., Chang-Claude, J.*, et al.* (1998). Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. American journal of human genetics *62*, 676-689.

Freedman, M.L., Monteiro, A.N., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D.*, et al.* (2011). Principles for the post-GWAS functional characterization of cancer risk loci. Nature genetics *43*, 513-518.

Fu, X.D., and Maniatis, T. (1990). Factor required for mammalian spliceosome assembly is localized to discrete regions in the nucleus. Nature *343*, 437-441.
Fuda, N.J., Ardehali, M.B., and Lis, J.T. (2009). Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature *461*, 186-192.

Fujita, M., Enomoto, T., Yoshino, K., Nomura, T., Buzard, G.S., Inoue, M., and Okudaira, Y. (1995). Microsatellite instability and alterations in the hMSH2 gene in human ovarian cancer. International journal of cancer Journal international du cancer *64*, 361-366.

Fulton, D.L., Sundararajan, S., Badis, G., Hughes, T.R., Wasserman, W.W., Roach, J.C., and Sladek, R. (2009). TFCat: the curated catalog of mouse and human transcription factors. Genome biology *10*, R29.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M.*, et al.* (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225-2229.

Gaszner, M., and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. Nature reviews Genetics *7*, 703-713.

Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.

Gerasimova, T.I., Byrd, K., and Corces, V.G. (2000). A chromatin insulator determines the nuclear localization of DNA. Molecular cell *6*, 1025-1035.

Gerstein, M.B., Kundaje, A., Hariharan, M., Landt, S.G., Yan, K.K., Cheng, C., Mu, X.J., Khurana, E., Rozowsky, J., Alexander, R.*, et al.* (2012). Architecture of the human regulatory network derived from ENCODE data. Nature *489*, 91-100.

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96-100.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome research *17*, 877-885.

Gomes, N.P., Bjerke, G., Llorente, B., Szostek, S.A., Emerson, B.M., and Espinosa, J.M. (2006). Gene-specific requirement for P-TEFb activity and RNA polymerase II phosphorylation within the p53 transcriptional program. Genes & development *20*, 601-612.

Gondor, A., and Ohlsson, R. (2009). Chromosome crosstalk in three dimensions. Nature *461*, 212-217.

Goode, E.L., Chenevix-Trench, G., Song, H., Ramus, S.J., Notaridou, M., Lawrenson, K., Widschwendter, M., Vierkant, R.A., Larson, M.C., Kjaer, S.K.*, et al.* (2010). A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. Nature genetics *42*, 874-879.

Group, E.C.W. (2005). Noncontraceptive health benefits of combined oral contraception. Human reproduction update *11*, 513-525.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W.*, et al.* (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948-951.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77-88.

Gunther, K., Rust, M., Leers, J., Boettger, T., Scharfe, M., Jarek, M., Bartkuhn, M., and Renkawitz, R. (2013). Differential roles for MBD2 and MBD3 at methylated CpG islands, active promoters and binding to exon sequences. Nucleic Acids Res *41*, 3010-3021.

Hagege, H., Klous, P., Braem, C., Splinter, E., Dekker, J., Cathala, G., de Laat, W., and Forne, T. (2007). Quantitative analysis of chromosome conformation capture assays (3C-qPCR). Nat Protoc *2*, 1722-1733.

Hankinson, S.E., Colditz, G.A., Hunter, D.J., Spencer, T.L., Rosner, B., and Stampfer, M.J. (1992). A quantitative assessment of oral contraceptive use and risk of ovarian cancer. Obstetrics and gynecology *80*, 708-714.

Hanna-Rose, W., and Hansen, U. (1996). Active repression mechanisms of eukaryotic transcription repressors. Trends in genetics : TIG *12*, 229-234.

Hanna, L., and Adams, M. (2006). Prevention of ovarian cancer. Best practice & research Clinical obstetrics & gynaecology *20*, 339-362.

Hardy, J., and Singleton, A. (2009). Genomewide association studies and human disease. The New England journal of medicine *360*, 1759-1768.

Hazelett, D.J., Rhie, S.K., Gaddis, M., Yan, C., Lakeland, D.L., Coetzee, S.G., Ellipse, G.-O.N.c., Practical, c., Henderson, B.E., Noushmehr, H.*, et al.* (2014). Comprehensive functional annotation of 77 prostate cancer risk loci. PLoS genetics *10*, e1004102.

Hebbar, P.B., and Archer, T.K. (2003). Chromatin remodeling by nuclear receptors. Chromosoma *111*, 495-504.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W.*, et al.* (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108-112.

Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A.*, et al.* (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nature genetics *39*, 311-318.

Hider, J.L., Gittelman, R.M., Shah, T., Edwards, M., Rosenbloom, A., Akey, J.M., and Parra, E.J. (2013). Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian ancestry. BMC evolutionary biology *13*, 150.

Hinkula, M., Pukkala, E., Kyyronen, P., and Kauppila, A. (2006). Incidence of ovarian cancer of grand multiparous women--a population-based study in Finland. Gynecologic oncology *103*, 207-211.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nature reviews Genetics *6*, 95-108.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-Andre, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934-947.

Hu, G., and Wade, P.A. (2012). NuRD and pluripotency: a complex balancing act. Cell Stem Cell *10*, 497-503.

Hunter, D.J., and Kraft, P. (2007). Drinking from the fire hose--statistical issues in genomewide association studies. The New England journal of medicine *357*, 436-439.
International HapMap, C. (2005). A haplotype map of the human genome. Nature *437*, 1299-1320.

Ioannidis, J.P., Thomas, G., and Daly, M.J. (2009). Validating, augmenting and refining genome-wide association signals. Nature reviews Genetics *10*, 318-329.

Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. Proceedings of the National Academy of Sciences of the United States of America *94*, 5617-5621.

Iuchi, S., and Green, H. (1999). Basonuclin, a zinc finger protein of keratinocytes and reproductive germ cells, binds to the rRNA gene promoter. Proceedings of the National Academy of Sciences of the United States of America *96*, 9628-9632.

Jacobs, L.C., Wollstein, A., Lao, O., Hofman, A., Klaver, C.C., Uitterlinden, A.G., Nijsten, T., Kayser, M., and Liu, F. (2013). Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. Human genetics *132*, 147-158.

Jakobsdottir, J., Gorin, M.B., Conley, Y.P., Ferrell, R.E., and Weeks, D.E. (2009). Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS genetics *5*, e1000337.

Jimenez-Sanchez, G., Childs, B., and Valle, D. (2001). Human disease genes. Nature *409*, 853-855.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.A., Schmitt, A.D., Espinoza, C.A., and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290-294.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S.*, et al.* (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430-435.

Kanchi, K.L., Johnson, K.J., Lu, C., McLellan, M.D., Leiserson, M.D., Wendl, M.C., Zhang, Q., Koboldt, D.C., Xie, M., Kandoth, C.*, et al.* (2014). Integrated analysis of germline and somatic variants in ovarian cancer. Nature communications *5*, 3156.
Kataoka, N., Yong, J., Kim, V.N., Velazquez, F., Perkinson, R.A., Wang, F., and Dreyfuss, G. (2000). Pre-mRNA splicing imprints mRNA in the nucleus with a novel RNA-binding protein that persists in the cytoplasm. Molecular cell *6*, 673-682.

Keaton, M.A., Taylor, C.M., Layer, R.M., and Dutta, A. (2011). Nuclear scaffold attachment sites within ENCODE regions associate with actively transcribed genes. PloS one *6*, e17912.

Kelemen, L.E., and Kobel, M. (2011). Mucinous carcinomas of the ovary and colorectum: different organ, same dilemma. The Lancet Oncology *12*, 1071-1080.

Keller, A., Nesvizhskii, A.I., Kolker, E., and Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem *74*, 5383-5392.

Kellum, R., and Schedl, P. (1992). A group of scs elements function as domain boundaries in an enhancer-blocking assay. Molecular and cellular biology *12*, 2424-2431.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome research *12*, 996-1006.

Kerr, L.D. (1995). Electrophoretic mobility shift assay. Methods in enzymology *254*, 619-632.

Khoury, G., and Gruss, P. (1983). Enhancer elements. Cell *33*, 313-314.
Kieffer-Kwon, K.R., Tang, Z., Mathe, E., Qian, J., Sung, M.H., Li, G., Resch, W., Baek, S., Pruett, N., Grontved, L.*, et al.* (2013). Interactome maps of mouse gene regulatory domains reveal basic principles of transcriptional regulation. Cell *155*, 1507-1520.

King, B.L., Carcangiu, M.L., Carter, D., Kiechle, M., Pfisterer, J., Pfleiderer, A., and Kacinski, B.M. (1995). Microsatellite instability in ovarian neoplasms. British journal of cancer *72*, 376-382.

Kisseljova, N.P., Dmitriev, P., Katargin, A., Kim, E., Ezerina, D., Markozashvili, D., Malysheva, D., Planche, E., Lemmers, R.J., van der Maarel, S.M.*, et al.* (2014). DNA polymorphism and epigenetic marks modulate the affinity of a scaffold/matrix attachment region to the nuclear matrix. European journal of human genetics : EJHG *22*, 1117-1123.

Klug, A. (2010). The discovery of zinc fingers and their applications in gene regulation and genome manipulation. Annual review of biochemistry *79*, 213-231.

Krause, S., Fakan, S., Weis, K., and Wahle, E. (1994). Immunodetection of poly(A) binding protein II in the cell nucleus. Experimental cell research *214*, 75-82.

Kuchenbaecker, K.B., Ramus, S.J., Tyrer, J., Lee, A., Shen, H.C., Beesley, J., Lawrenson, K., McGuffog, L., Healey, S., Lee, J.M.*, et al.* (2015). Identification of six new susceptibility loci for invasive epithelial ovarian cancer. Nature genetics.

Kulaeva, O.I., Nizovtseva, E.V., Polikanov, Y.S., Ulianov, S.V., and Studitsky, V.M. (2012). Distant activation of transcription: mechanisms of enhancer action. Molecular and cellular biology *32*, 4892-4897.

Kuusisto, K.M., Bebel, A., Vihinen, M., Schleutker, J., and Sallinen, S.L. (2011). Screening for BRCA1, BRCA2, CHEK2, PALB2, BRIP1, RAD50, and CDH1 mutations in high-risk Finnish BRCA1/2-founder mutation-negative breast and/or ovarian cancer individuals. Breast cancer research : BCR *13*, R20.

Lam, K.N., van Bakel, H., Cote, A.G., van der Ven, A., and Hughes, T.R. (2011). Sequence specificity is obtained from the majority of modular C2H2 zinc-finger arrays. Nucleic acids research *39*, 4680-4690.

Lamond, A.I., and Spector, D.L. (2003). Nuclear speckles: a model for nuclear organelles. Nature reviews Molecular cell biology *4*, 605-612.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lang, M.R., Patterson, L.B., Gordon, T.N., Johnson, S.L., and Parichy, D.M. (2009). Basonuclin-2 requirements for zebrafish adult pigment pattern development and female fertility. PLoS genetics *5*, e1000744.

Lawrenson, K., Benjamin, E., Turmaine, M., Jacobs, I., Gayther, S., and Dafou, D. (2009). In vitro three-dimensional modelling of human ovarian surface epithelial cells. Cell Prolif *42*, 385-393.

Lawrenson, K., Grun, B., Benjamin, E., Jacobs, I.J., Dafou, D., and Gayther, S.A. (2010). Senescent fibroblasts promote neoplastic transformation of partially transformed ovarian epithelial cells in a three-dimensional model of early stage ovarian cancer. Neoplasia *12*, 317-325.

Lee, K.R., and Young, R.H. (2003). The distinction between primary and metastatic mucinous carcinomas of the ovary: gross and histologic findings in 50 cases. The American journal of surgical pathology *27*, 281-292.

Lee, T.I., and Young, R.A. (2000). Transcription of eukaryotic protein-coding genes. Annual review of genetics *34*, 77-137.

Lee, Y., Miron, A., Drapkin, R., Nucci, M.R., Medeiros, F., Saleemuddin, A., Garber, J., Birch, C., Mou, H., Gordon, R.W.*, et al.* (2007). A candidate precursor to serous carcinoma that originates in the distal fallopian tube. The Journal of pathology *211*, 26-35.

Leong, H.S., Galletta, L., Etemadmoghadam, D., George, J., Australian Ovarian Cancer, S., Kobel, M., Ramus, S.J., and Bowtell, D. (2015). Efficient molecular subtype classification of high-grade serous ovarian cancer. The Journal of pathology *236*, 272-277.

Li, Q., Harju, S., and Peterson, K.R. (1999). Locus control regions: coming of age at a decade plus. Trends in genetics : TIG *15*, 403-408.

Li, Q., Seo, J.H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T., Brown, M., Tyekucheva, S., and Freedman, M.L. (2013). Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell *152*, 633-641.

Lindblom, B., Hamberger, L., and Ljung, B. (1980). Contractile patterns of isolated oviductal smooth muscle under different hormonal conditions. Fertility and sterility *33*, 283-287.

Linnemann, A.K., Platts, A.E., and Krawetz, S.A. (2009). Differential nuclear scaffold/matrix attachment marks expressed genes. Human molecular genetics *18*, 645-654.

Liu, T., Ortiz, J.A., Taing, L., Meyer, C.A., Lee, B., Zhang, Y., Shin, H., Wong, S.S., Ma, J., Lei, Y.*, et al.* (2011). Cistrome: an integrative platform for transcriptional regulation studies. Genome biology *12*, R83.

Lu, K.H., and Daniels, M. (2013). Endometrial and ovarian cancer in women with Lynch syndrome: update in screening and prevention. Familial cancer *12*, 273-277.

Machanick, P., and Bailey, T.L. (2011). MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics *27*, 1696-1697.

Mahoney, M.G., Tang, W., Xiang, M.M., Moss, S.B., Gerton, G.L., Stanley, J.R., and Tseng, H. (1998). Translocation of the zinc finger protein basonuclin from the mouse germ cell nucleus to the midpiece of the spermatozoon during spermiogenesis. Biology of reproduction *59*, 388-394.

Maisey, K., Nardocci, G., Imarai, M., Cardenas, H., Rios, M., Croxatto, H.B., Heckels, J.E., Christodoulides, M., and Velasquez, L.A. (2003). Expression of proinflammatory cytokines and receptors by human fallopian tubes in organ culture following challenge with Neisseria gonorrhoeae. Infection and immunity *71*, 527-532.

Malander, S., Rambech, E., Kristoffersson, U., Halvarsson, B., Ridderheim, M., Borg, A., and Nilbert, M. (2006). The contribution of the hereditary nonpolyposis colorectal cancer syndrome to the development of ovarian cancer. Gynecologic oncology *101*, 238-243.

Manolio, T.A. (2010). Genomewide association studies and assessment of the risk of disease. The New England journal of medicine *363*, 166-176.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A.*, et al.* (2009). Finding the missing heritability of complex diseases. Nature *461*, 747-753.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J.*, et al.* (2012). Systematic localization of common disease-associated variation in regulatory DNA. Science *337*, 1190-1195.

McGee, Z.A., Jensen, R.L., Clemens, C.M., Taylor-Robinson, D., Johnson, A.P., and Gregg, C.R. (1999). Gonococcal infection of human fallopian tube mucosa in organ culture: relationship of mucosal tissue TNF-alpha concentration to sloughing of ciliated cells. Sexually transmitted diseases *26*, 160-165.

Meindl, A., Hellebrand, H., Wiek, C., Erven, V., Wappenschmidt, B., Niederacher, D., Freund, M., Lichtner, P., Hartmann, L., Schaal, H.*, et al.* (2010). Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. Nature genetics *42*, 410-414.

Messina, D.N., Glasscock, J., Gish, W., and Lovett, M. (2004). An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. Genome research *14*, 2041-2047.

Meyer, L.A., Broaddus, R.R., and Lu, K.H. (2009). Endometrial cancer and Lynch syndrome: clinical and pathologic considerations. Cancer control : journal of the Moffitt Cancer Center *16*, 14-22.

Miki, Y., Swensen, J., Shattuck-Eidens, D., Futreal, P.A., Harshman, K., Tavtigian, S., Liu, Q., Cochran, C., Bennett, L.M., Ding, W.*, et al.* (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science *266*, 66-71.

Mikkelsen, T.S., Ku, M., Jaffe, D.B., Issac, B., Lieberman, E., Giannoukos, G., Alvarez, P., Brockman, W., Kim, T.K., Koche, R.P.*, et al.* (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*, 553-560.

Minion, L.E., Dolinsky, J.S., Chase, D.M., Dunlop, C.L., Chao, E.C., and Monk, B.J. (2015). Hereditary predisposition to ovarian cancer, looking beyond BRCA1/BRCA2. Gynecologic oncology *137*, 86-92.

Mirkovitch, J., Mirault, M.E., and Laemmli, U.K. (1984). Organization of the higher-order chromatin loop: specific DNA attachment sites on nuclear scaffold. Cell *39*, 223-232.

Modugno, F., Ness, R.B., Allen, G.O., Schildkraut, J.M., Davis, F.G., and Goodman, M.T. (2004). Oral contraceptive use, reproductive history, and risk of epithelial ovarian cancer in women with and without endometriosis. American journal of obstetrics and gynecology *191*, 733-740.

Monteiro, A.N., and Freedman, M.L. (2013). Lessons from postgenome-wide association studies: functional analysis of cancer predisposition loci. Journal of internal medicine *274*, 414-424.

Najafabadi, H.S., Mnaimneh, S., Schmitges, F.W., Garton, M., Lam, K.N., Yang, A., Albu, M., Weirauch, M.T., Radovani, E., Kim, P.M.*, et al.* (2015). C2H2 zinc finger proteins greatly expand the human regulatory lexicon. Nature biotechnology.

Narlikar, G.J., Fan, H.Y., and Kingston, R.E. (2002). Cooperation between complexes that regulate chromatin structure and transcription. Cell *108*, 475-487.

Nash, M.A., Ferrandina, G., Gordinier, M., Loercher, A., and Freedman, R.S. (1999). The role of cytokines in both the normal and malignant ovary. Endocrine-related cancer *6*, 93-107.

Nef, S., Schaad, O., Stallings, N.R., Cederroth, C.R., Pitetti, J.L., Schaer, G., Malki, S., Dubois-Dauphin, M., Boizet-Bonhoure, B., Descombes, P.*, et al.* (2005). Gene expression during sex determination reveals a robust female genetic program at the onset of ovarian development. Dev Biol *287*, 361-377.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K.*, et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature *489*, 83-90.

Ness, R.B., Grisso, J.A., Cottreau, C., Klapper, J., Vergona, R., Wheeler, J.E., Morgan, M., and Schlesselman, J.J. (2000). Factors related to inflammation of the ovarian epithelium and risk of ovarian cancer. Epidemiology *11*, 111-117.

Nesvizhskii, A.I., Keller, A., Kolker, E., and Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem *75*, 4646-4658.

Nezhat, F., Datta, M.S., Hanson, V., Pejovic, T., Nezhat, C., and Nezhat, C. (2008). The relationship of endometriosis and ovarian malignancy: a review. Fertility and sterility *90*, 1559-1570.

Ohlsson, R., Renkawitz, R., and Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. Trends in genetics : TIG *17*, 520-527.

Pagano, J.S., Blaser, M., Buendia, M.A., Damania, B., Khalili, K., Raab-Traub, N., and Roizman, B. (2004). Infectious agents and cancer: criteria for a causal relation. Seminars in cancer biology *14*, 453-471.

Pal, T., and Bhattacharyya, A.K. (1989). Structural changes in human cervical mucus. The Indian journal of medical research *90*, 44-50.

Pascuzzi, P.E., Flores-Vergara, M.A., Lee, T.J., Sosinski, B., Vaughn, M.W., Hanley-Bowdoin, L., Thompson, W.F., and Allen, G.C. (2014). In vivo mapping of arabidopsis scaffold/matrix attachment regions reveals link to nucleosome-disfavoring poly(dA:dT) tracts. The Plant cell *26*, 102-120.

Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809-817.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D*., et al.* (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499-502.

Pennington, K.P., and Swisher, E.M. (2012). Hereditary ovarian cancer: beyond the usual suspects. Gynecologic oncology *124*, 347-353.

Pennington, K.P., Walsh, T., Harrell, M.I., Lee, M.K., Pennil, C.C., Rendi, M.H., Thornton, A., Norquist, B.M., Casadei, S., Nord, A.S*., et al.* (2014). Germline and somatic mutations in homologous recombination genes predict platinum response and survival in ovarian, fallopian tube, and peritoneal carcinomas. Clinical cancer research : an official journal of the American Association for Cancer Research *20*, 764-775.

Perkins, D.N., Pappin, D.J., Creasy, D.M., and Cottrell, J.S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis *20*, 3551-3567.

Permuth-Wey, J., Lawrenson, K., Shen, H.C., Velkova, A., Tyrer, J.P., Chen, Z., Lin, H.Y., Ann Chen, Y., Tsai, Y.Y., Qu, X.*, et al.* (2013). Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31. Nature communications *4*, 1627.

Pharoah, P.D., Dunning, A.M., Ponder, B.A., and Easton, D.F. (2004). Association studies for finding cancer-susceptibility genetic variants. Nature reviews Cancer *4*, 850-860.

Pharoah, P.D., and Ponder, B.A. (2002). The genetics of ovarian cancer. Best practice & research Clinical obstetrics & gynaecology *16*, 449-468.

Pharoah, P.D., Tsai, Y.Y., Ramus, S.J., Phelan, C.M., Goode, E.L., Lawrenson, K., Buckley, M., Fridley, B.L., Tyrer, J.P., Shen, H.*, et al.* (2013). GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nature genetics *45*, 362-370.

Piek, J.M., van Diest, P.J., Zweemer, R.P., Jansen, J.W., Poort-Keesom, R.J., Menko, F.H., Gille, J.J., Jongsma, A.P., Pals, G., Kenemans, P.*, et al.* (2001). Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer. The Journal of pathology *195*, 451-456.

Plank, J.L., and Dean, A. (2014). Enhancer function: mechanistic and genome-wide insights come together. Molecular cell *55*, 5-14.

Pritchard, J.K. (2001). Are rare variants responsible for susceptibility to complex diseases? American journal of human genetics *69*, 124-137.

Rafnar, T., Gudbjartsson, D.F., Sulem, P., Jonasdottir, A., Sigurdsson, A., Jonasdottir, A., Besenbacher, S., Lundin, P., Stacey, S.N., Gudmundsson, J.*, et al.* (2011). Mutations in BRIP1 confer high risk of ovarian cancer. Nature genetics *43*, 1104-1107.

Ramakrishna, M., Williams, L.H., Boyle, S.E., Bearfoot, J.L., Sridhar, A., Speed, T.P., Gorringe, K.L., and Campbell, I.G. (2010). Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. PLoS One *5*, e9983.

Ramirez, C.L., Foley, J.E., Wright, D.A., Muller-Lerch, F., Rahman, S.H., Cornu, T.I., Winfrey, R.J., Sander, J.D., Fu, F., Townsend, J.A.*, et al.* (2008). Unexpected failure rates for modular assembly of engineered zinc fingers. Nat Methods *5*, 374-375.

Ramirez, J., and Hagman, J. (2009). The Mi-2/NuRD complex: a critical epigenetic regulator of hematopoietic development, differentiation and cancer. Epigenetics *4*, 532-536.

Ramus, S.J., Antoniou, A.C., Kuchenbaecker, K.B., Soucy, P., Beesley, J., Chen, X., McGuffog, L., Sinilnikova, O.M., Healey, S., Barrowdale, D*., et al.* (2012). Ovarian cancer susceptibility alleles and risk of ovarian cancer in BRCA1 and BRCA2 mutation carriers. Human mutation *33*, 690-702.

Ramus, S.J., Harrington, P.A., Pye, C., DiCioccio, R.A., Cox, M.J., Garlinghouse-Jones, K., Oakley-Girvan, I., Jacobs, I.J., Hardy, R.M., Whittemore, A.S*., et al.* (2007). Contribution of BRCA1 and BRCA2 mutations to inherited ovarian cancer. Human mutation *28*, 1207-1215.

Ramus, S.J., Kartsonaki, C., Gayther, S.A., Pharoah, P.D., Sinilnikova, O.M., Beesley, J., Chen, X., McGuffog, L., Healey, S., Couch, F.J*., et al.* (2011). Genetic variation at 9p22.2 and ovarian cancer risk for BRCA1 and BRCA2 mutation carriers. Journal of the National Cancer Institute *103*, 105-116.

Ravasi, T., Huber, T., Zavolan, M., Forrest, A., Gaasterland, T., Grimmond, S., Hume, D.A., Group, R.G., and Members, G.S.L. (2003). Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. Genome research *13*, 1430-1442.

Reich, D.E., and Lander, E.S. (2001). On the allelic spectrum of human disease. Trends in genetics : TIG *17*, 502-510.

Riman, T., Persson, I., and Nilsson, S. (1998). Hormonal aspects of epithelial ovarian cancer: review of epidemiological evidence. Clinical endocrinology *49*, 695-707.

Risch, H.A. (1998). Hormonal etiology of epithelial ovarian cancer, with a hypothesis concerning the role of androgens and progesterone. Journal of the National Cancer Institute *90*, 1774-1786.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516-1517.

Risch, N.J. (2000). Searching for genetic determinants in the new millennium. Nature *405*, 847-856.

Romano, R.A., Li, H., Tummala, R., Maul, R., and Sinha, S. (2004). Identification of Basonuclin2, a DNA-binding zinc-finger protein expressed in germ tissues and skin keratinocytes. Genomics *83*, 821-833.

Salehi, F., Dunfield, L., Phillips, K.P., Krewski, D., and Vanderhyden, B.C. (2008). Risk factors for ovarian cancer: an overview with emphasis on hormonal factors. Journal of toxicology and environmental health Part B, Critical reviews *11*, 301-321.

Salvador, S., Gilks, B., Kobel, M., Huntsman, D., Rosen, B., and Miller, D. (2009). The fallopian tube: primary site of most pelvic high-grade serous carcinomas. International journal of gynecological cancer : official journal of the International Gynecological Cancer Society *19*, 58-64.

Sankararaman, S., Mallick, S., Dannemann, M., Prufer, K., Kelso, J., Paabo, S., Patterson, N., and Reich, D. (2014). The genomic landscape of Neanderthal ancestry in present-day humans. Nature *507*, 354-357.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109-113.
Schindler, R., Nilsson, E., and Skinner, M.K. (2010). Induction of ovarian primordial follicle assembly by connective tissue growth factor CTGF. PLoS One *5*, e12979.

Shen, H., Fridley, B.L., Song, H., Lawrenson, K., Cunningham, J.M., Ramus, S.J., Cicek, M.S., Tyrer, J., Stram, D., Larson, M.C*., et al.* (2013). Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. Nature communications *4*, 1628.

Song, H., Ramus, S.J., Tyrer, J., Bolton, K.L., Gentry-Maharaj, A., Wozniak, E., Anton-Culver, H., Chang-Claude, J., Cramer, D.W., DiCioccio, R*., et al.* (2009). A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. Nature genetics *41*, 996-1000.

Strandell, A., Thorburn, J., and Wallin, A. (2004). The presence of cytokines and growth factors in hydrosalpingeal fluid. Journal of assisted reproduction and genetics *21*, 241-247.

Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. Genetics *187*, 367-383.
Stratton, J.F., Pharoah, P., Smith, S.K., Easton, D., and Ponder, B.A. (1998). A systematic review and meta-analysis of family history and risk of ovarian cancer. British journal of obstetrics and gynaecology *105*, 493-499.

Struhl, K. (1998). Histone acetylation and transcriptional regulatory mechanisms. Genes & development *12*, 599-606.
Sueblinvong, T., and Carney, M.E. (2009). Current understanding of risk factors for ovarian cancer. Current treatment options in oncology *10*, 67-81.

Swift, S., Lorens, J., Achacoso, P., and Nolan, G.P. (2001). Rapid production of retroviruses for efficient gene delivery to mammalian cells using 293T cell-based systems. Curr Protoc Immunol *Chapter 10*, Unit 10 17C.

Tan, T.Z., Miow, Q.H., Huang, R.Y., Wong, M.K., Ye, J., Lau, J.A., Wu, M.C., Bin Abdul Hadi, L.H., Soong, R., Choolani, M.*, et al.* (2013). Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. EMBO molecular medicine *5*, 983-998.

TCGA (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609-615.
Thorstenson, Y.R., Roxas, A., Kroiss, R., Jenkins, M.A., Yu, K.M., Bachrich, T., Muhr, D., Wayne, T.L., Chu, G., Davis, R.W.*, et al.* (2003). Contributions of ATM mutations to familial breast and ovarian cancer. Cancer research *63*, 3325-3333.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B.*, et al.* (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75-82.

Tian, Q., Kopf, G.S., Brown, R.S., and Tseng, H. (2001). Function of basonuclin in increasing transcription of the ribosomal RNA genes during mouse oogenesis. Development *128*, 407-416.

Tothill, R.W., Tinker, A.V., George, J., Brown, R., Fox, S.B., Lade, S., Johnson, D.S., Trivett, M.K., Etemadmoghadam, D., Locandro, B.*, et al.* (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clinical cancer research : an official journal of the American Association for Cancer Research *14*, 5198-5208.

Tseng, H., Biegel, J.A., and Brown, R.S. (1999). Basonuclin is associated with the ribosomal RNA genes on human keratinocyte mitotic chromosomes. Journal of cell science *112 Pt 18*, 3039-3047.

Tseng, H., and Green, H. (1992). Basonuclin: a keratinocyte protein with multiple paired zinc fingers. Proceedings of the National Academy of Sciences of the United States of America *89*, 10311-10315.

Tupler, R., Perini, G., and Green, M.R. (2001). Expressing the human genome. Nature *409*, 832-833.

Turner, N., Tutt, A., and Ashworth, A. (2004). Hallmarks of 'BRCAness' in sporadic cancers. Nature reviews Cancer *4*, 814-819.
van Arensbergen, J., van Steensel, B., and Bussemaker, H.J. (2014). In search of the determinants of enhancer-promoter interaction specificity. Trends in cell biology *24*, 695-702.

Vanhoutteghem, A., Bouche, C., Maciejewski-Duval, A., Herve, F., and Djian, P. (2011). Basonuclins and disco: Orthologous zinc finger proteins essential for development in vertebrates and arthropods. Biochimie *93*, 127-133.

Vanhoutteghem, A., and Djian, P. (2004). Basonuclin 2: an extremely conserved homolog of the zinc finger protein basonuclin. Proceedings of the National Academy of Sciences of the United States of America *101*, 3468-3473.

Vanhoutteghem, A., and Djian, P. (2006). Basonuclins 1 and 2, whose genes share a common origin, are proteins with widely different properties and functions. Proceedings of the National Academy of Sciences of the United States of America *103*, 12423-12428.

Vanhoutteghem, A., and Djian, P. (2007). The human basonuclin 2 gene has the potential to generate nearly 90,000 mRNA isoforms encoding over 2000 different proteins. Genomics *89*, 44-58.

Vanhoutteghem, A., Maciejewski-Duval, A., Bouche, C., Delhomme, B., Herve, F., Daubigney, F., Soubigou, G., Araki, M., Araki, K., Yamamura, K.*, et al.* (2009). Basonuclin 2 has a function in the multiplication of embryonic craniofacial mesenchymal cells and is orthologous to disco proteins. Proc Natl Acad Sci U S A *106*, 14432-14437.

Vanhoutteghem, A., Messiaen, S., Herve, F., Delhomme, B., Moison, D., Petit, J.M., Rouiller-Fabre, V., Livera, G., and Djian, P. (2014). The zinc-finger protein basonuclin 2 is required for proper mitotic arrest, prevention of premature meiotic initiation and meiotic progression in mouse male germ cells. Development *141*, 4298-4310.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., and Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. Nature reviews Genetics *10*, 252-263.

Vaughan, S., Coward, J.I., Bast, R.C., Jr., Berchuck, A., Berek, J.S., Brenton, J.D., Coukos, G., Crum, C.C., Drapkin, R., Etemadmoghadam, D.*, et al.* (2011). Rethinking ovarian cancer: recommendations for improving outcomes. Nature reviews Cancer *11*, 719-725.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A.*, et al.* (2001). The sequence of the human genome. Science *291*, 1304-1351.

Vernot, B., and Akey, J.M. (2014). Resurrecting surviving Neandertal lineages from modern human genomes. Science *343*, 1017-1021.

Visser, M., Palstra, R.J., and Kayser, M. (2014). Human skin color is influenced by an intergenic DNA polymorphism regulating transcription of the nearby BNC2 pigmentation gene. Hum Mol Genet.

Walsh, T., Casadei, S., Lee, M.K., Pennil, C.C., Nord, A.S., Thornton, A.M., Roeb, W., Agnew, K.J., Stray, S.M., Wickramanayake, A.*, et al.* (2011). Mutations in 12 genes for inherited ovarian, fallopian tube, and peritoneal carcinoma identified by massively parallel sequencing. Proceedings of the National Academy of Sciences of the United States of America *108*, 18032-18037.

Weirauch, M.T., and Hughes, T.R. (2011). A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. Sub-cellular biochemistry *52*, 25-73.

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K.*, et al.* (2014). Determination and inference of eukaryotic transcription factor sequence specificity. Cell *158*, 1431-1443.

West, A.G., Gaszner, M., and Felsenfeld, G. (2002). Insulators: many functions, many mechanisms. Genes & development *16*, 271-288.

Wolfe, S.A., Grant, R.A., Elrod-Erickson, M., and Pabo, C.O. (2001). Beyond the "recognition code": structures of two Cys2His2 zinc finger/TATA box complexes. Structure *9*, 717-723.

Wolfe, S.A., Nekludova, L., and Pabo, C.O. (2000). DNA recognition by Cys2His2 zinc finger proteins. Annual review of biophysics and biomolecular structure *29*, 183-212.

Wood, A.R., Esko, T., Yang, J., Vedantam, S., Pers, T.H., Gustafsson, S., Chu, A.Y., Estrada, K., Luan, J., Kutalik, Z.*, et al.* (2014). Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet *46*, 1173-1186.

Woods, N.T., Mesquita, R.D., Sweet, M., Carvalho, M.A., Li, X., Liu, Y., Nguyen, H., Thomas, C.E., Iversen, E.S., Jr., Marsillac, S.*, et al.* (2012). Charting the landscape of tandem BRCT domain-mediated protein interactions. Sci Signal *5*, rs6.

Wooster, R., Bignell, G., Lancaster, J., Swift, S., Seal, S., Mangion, J., Collins, N., Gregory, S., Gumbs, C., and Micklem, G. (1995). Identification of the breast cancer susceptibility gene BRCA2. Nature *378*, 789-792.

Workman, J.L., and Kingston, R.E. (1998). Alteration of nucleosome structure as a mechanism of transcriptional regulation. Annual review of biochemistry *67*, 545-579.

Wray, N.R., Goddard, M.E., and Visscher, P.M. (2008). Prediction of individual genetic risk of complex disease. Current opinion in genetics & development *18*, 257-263.

Yang, W.M., Yao, Y.L., and Seto, E. (2001). The FK506-binding protein 25 functionally associates with histone deacetylases and with transcription factor YY1. The EMBO journal *20*, 4814-4825.

Yang, Z., Gallicano, G.I., Yu, Q.C., and Fuchs, E. (1997). An unexpected localization of basonuclin in the centrosome, mitochondria, and acrosome of developing spermatids. The Journal of cell biology *137*, 657-669.

Zhang, B., Kirov, S., and Snoddy, J. (2005). WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic acids research *33*, W741-748.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W.*, et al.* (2008). Model-based analysis of ChIP-Seq (MACS). Genome biology *9*, R137.

Zhang, Y., Wong, C.H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E.*, et al.* (2013). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature *504*, 306-310.

Zheng, H., Kavanagh, J.J., Hu, W., Liao, Q., and Fu, S. (2007). Hormonal therapy in ovarian cancer. International journal of gynecological cancer : official journal of the International Gynecological Cancer Society *17*, 325-338.

Zhou, Z., Luo, M.J., Straesser, K., Katahira, J., Hurt, E., and Reed, R. (2000). The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. Nature *407*, 401-405.

ABOUT THE AUTHOR

Melissa Ann Buckley (Price) grew up in Arvada, Colorado hiking, skiing, and playing soccer. She attended the University of Colorado in Boulder for undergraduate and as the first in her family to attend a state university. From day one she wanted to study the sciences and was a Molecular Cellular and Developmental Biology major.

Melissa's lab experience first began during an internship with a small pharmaceutical company, Replidyne in Louisville, CO. For her main project she measured the changes in activity of mutant methionly-tRNA synthetase of *C. dificile*. She thoroughly enjoyed the collaborative and creative atmosphere. She knew after this experience she wanted to continue her research career, thus she applied to the Cancer Biology PhD program at Moffitt Cancer Center, University of South Florida.

Her recent research was conducted in the lab of Dr. Alvaro Monteiro. Her thesis project included identifying the regulatory regions that overlap with SNPs identified in genome wide association studies for ovarian cancer and identifying the downstream target genes of those regulatory regions. Broadly, the Monteiro lab took on the research of several potential regulatory regions for ovarian cancer predisposition. Melissa helped develop many of the assays not previously performed in the lab but required for the emerging field; truly a challenge.

While in the Monteiro lab Melissa received an award from the ARCS (Achievement Reward for College Scientists) Foundation. She also obtained the Ruth L. Kirschstein National Research Service Award from the NIH.