Capture Hi-C to identify regulatory variants and target genes influencing breast cancer risk

---

Alisa Zvereva

The Institute of Cancer Research

University of London

This thesis is submitted for the degree of Doctor of Philosophy

August 2022

Statement of Contribution

I declare that I designed and carried out all the work presented in this thesis with the following exceptions:

Correlated variants required for the target enrichment array design (Section 2.5) were selected by Andrea Gillespie.

Harriet Kemp, Andrea Gillespie and Syed Haider carried out the bioinformatics required to convert the raw sequencing data from the CHi-C experiments into a set of statistically significant interaction peaks. The downstream analysis of the interaction peaks was carried out by me, with the exception of the overlapping of the interaction peaks with annotated gene promoters and CCVs.

Harriet Kemp and Syed Haider also helped to generate the subset of figures in which looping interactions were aligned with other relevant features including RefSeq genes, capture baits, H3K27ac peaks and CTCF ChIP-seq data.

Harriet Kemp carried out the analysis to identify whether the third-party bins are enriched for CTCF and H3K27ac binding in Section 4.3.

CUT&Tag H3K27ac data and RNA-seq data were generated by other members of the Functional Genetic Epidemiology lab.

Signature: Alisa Zvereva

Abstract

Genome-wide association studies have discovered approximately 200 breast cancer risk single nucleotide polymorphisms, most of which map to non-protein-coding regions. To understand the mechanisms influencing disease risk, identification of the genes, non-coding RNAs and causal variants mediating these associations is required. One of the methods that allows functional characterisation of cancer risk loci is Capture Hi-C (CHi-C). CHi-C provides a high-throughput, high-resolution approach for studying physical interactions between long-range regulatory elements and their targets and has previously been used to identify putative target genes and to prioritise credible variants at a subset of risk loci. To date, however, CHi-C data have only been generated in breast cancer and immortalised 'normal' breast epithelial cell lines. Additionally, most studies have used HindIII digested libraries, which result in an average resolution of 10 kb.

The aims of this project were to:

1. Generate region CHi-C libraries in breast epithelial and fibroblast cell lines using three different protocols to identify and optimise the most suitable method for library generation in primary cells;
2. Generate higher resolution region CHi-C data in two types of primary breast cells (luminal epithelial cells and fibroblasts) to identify regulatory variants and target genes influencing breast cancer risk;
3. Compare cell line data to the primary cell data to evaluate the usefulness of cell lines as model systems;
4. Generate cell line promoter CHi-C data to validate region CHi-C findings and to identify 'indirect' interactions;
5. Using data generated in primary breast fibroblasts, determine whether a subset of breast cancer loci may act via the stroma to influence the risk.

Statement of Impact

The COVID-19 pandemic has inevitably impacted my PhD. Specifically, during the first lockdown period, the Institute of Cancer Research had closed its laboratories, so I was unable to carry out lab work between 25/03/2020 and 22/06/2020. Upon return on-site, we had to work in shifts for approximately three months which also limited the amount of lab work I could have done. Sequencing of some of my CHi-C libraries that were sequenced at the ICR Genomics Facility was delayed for the same reasons. Most companies operated reduced services during the pandemic, therefore, deliveries of some reagents and provision of some services (e.g., generation of functional datasets) were delayed. Due to limited number of staff in the Breast Cancer Now Tissue Bank, delivery of primary cells that I used for my rCHi-C libraries was delayed as well. All of the above delayed the generation of some of my CHi-C libraries. Despite the aforementioned disruptions, I have generated and analysed rCHi-C libraries generated using three different protocols in two different cell lines. Using my preferred methodology (Dovetail Genomics Omni-C kit), I have also generated pCHi-C libraries in the same two cell lines and rCHi-C libraries in two different primary cell types. These data formed the basis of my thesis.

As an overseas student, the restrictions around my visa have limited my options for extending the duration of my PhD. Given more time, I would have been able to systematically integrate my data with additional functional data generated by the Functional Genetic Epidemiology lab. I would have also been able to perform integration analysis (rCHi-C + pCHi-C) in primary breast luminal epithelial cells and fibroblasts (instead of just T-47D and GS2 cell lines) and to carry out initial functional follow up of one or more ‘proof of principle’ risk loci (using reporter gene assay and CRISPR).

The rCHi-C and pCHi-C data generated in T-47D and GS2 cell lines are the focus of a manuscript that is at an advanced stage of preparation and should be submitted in the next few months.

Candidate signature: Alisa Zvereva

Supervisor signature: Olivia Fletcher

Table of Contents

List of Figures

List of Tables

1. Introduction

1.1. Breast Cancer

Breast cancer is currently the most commonly diagnosed cancer in the world, with an estimated number of 2.3 million new cases diagnosed in 2020, representing 11.7% of all cancer cases[1]. It is also the fifth leading cause of cancer-related deaths with 685 thousand deaths recorded in 2020 worldwide. Among women, breast cancer accounts for 1 in 4 cancer cases and 1 in 6 cancer-related deaths, ranking first for incidence in 159 of 185 countries and for mortality in 110 countries.

Incidence rates are 88% higher in transitioned versus transitioning countries (55.9 and 29.7 per 100,000, respectively), possibly reflecting a higher prevalence of reproductive and hormonal risk factors in the transitioned countries. However, mortality rates were found to be 17% higher in transitioning countries (15.0 and 12.8 per 100,000, respectively), potentially reflecting weak health infrastructure and hence late-stage diagnosis and poor survival outcomes.

1.1.1. Classification

Histological analysis of breast biopsies broadly divides breast cancers into non-invasive and invasive. Ductal carcinoma in situ (DCIS) and lobular carcinoma in situ (LCIS) are non-invasive premalignant lesions, which can become invasive if left untreated. Invasive breast cancers (IBCs) are, however, highly heterogeneous, with at least 18 different types being described by the World Health Organization (WHO)[2]. Invasive ductal carcinoma not otherwise specified (IDC NOS) is the most commonly diagnosed breast cancer, accounting for 50-80% of cases. IDC NOS is diagnosed by default when a tumour fails to display sufficient morphological characteristics to be assigned to one of the histological special types. Around 25% of IBCs display distinctive growth patterns and cytological features and are therefore recognised as ‘special types’. These include invasive lobular carcinoma (ILC) – the second most common subtype (5-15% of cases) that is defined by epithelial cadherin mutations and a dissociated growth pattern; and other less common types include tubular, medullary and neuroendocrine carcinomas (each < 7% of cases). Although histological analysis is a valuable tool, it does not consider newer molecular markers that have a proven prognostic significance and, therefore, does not allow precise

stratification of patients and treatment options. To overcome these limitations, molecular classification systems were developed.

The most well-established molecular classification of breast cancer subtypes relies on immunohistochemical (IHC) staining. This classification considers expression of three molecular markers in tissue sections: estrogen receptor alpha (ER-α, *ESR1*), progesterone receptor (PR, *PGR*) and human epidermal growth factor 2 receptor (HER2, *ERBB2*). Hormone receptor positive (ER+/PR+)/*ERBB2* negative cancers comprise ~70% of cases, *ERBB2* positive (ERBB2+) ~15-20% and triple-negative breast cancer (TNBC) is diagnosed in ~15% of patients[3]. Hormone receptor subtyping is simple, quick and cost-effective; however, it does not account for the true extent of variability between breast cancers.

In 2000, Perou and colleagues used microarray-based gene expression to define four molecular classes of breast cancers: luminal, HER2-enriched, basal-like and normal breast-like[4]. Further studies sub-divided luminal cancers into two distinct subgroups (luminal A and B)[5, 6] while normal breast-like subtype has been omitted, as it likely represented sample contamination by normal mammary glands. In The Cancer Genome Atlas (TCGA) project, profiling of > 300 tumours at DNA, RNA and protein levels confirmed four main breast cancer intrinsic subtypes – luminal A, luminal B, HER2-enriched and basal like[7]. Later, a fifth subtype was discovered – claudin-low breast cancer that is defined by the low expression of key components of cellular junctions, associated with mesenchymal and stemness features[8, 9].

In 2009, PAM50, a 50-gene signature for subtype assignment, was developed that allows breast cancer classification into the main intrinsic subtypes with 93% accuracy[10] and is now clinically implemented worldwide.

1.1.2. Risk factors

There are many established breast cancer risk factors which include both modifiable and non-modifiable factors.

1.1.2.1. Modifiable risk factors

Modifiable risk factors can also be referred to as lifestyle risk factors. These factors are of great interest to research, since widely adopted lifestyle changes could decrease breast cancer incidence at a population level. Multiple lifestyle factors influencing breast cancer predisposition have been identified to date.

First, lack of physical activity has been associated with increased breast cancer risk[11]. Although the exact mechanism remains unclear, several potential explanations for the protective role of exercising have been proposed, including reduced exposure to the endogenous sex hormones[12], altered immune system responses[13] or elevated insulin-like growth factor 1 (IGF-1) levels[14].

Increased Body Mass Index (BMI) has also been associated with a higher probability of developing breast cancer[15]. The association is the strongest in obese post-menopausal women who are at increased risk, specifically, of ER+ breast cancer. Independently of menopausal status, obesity has also been associated with poorer clinical outcomes. Although BMI is a useful measure, it worth noting that it neither distinguishes lean mass from fat mass, nor characterises body fat distribution, and so individuals with the same BMI can have different body composition. As a result, further understanding of what aspects of body composition are the most important in determining risk is required. Some studies propose that higher body fat might lead to increased inflammation and affect levels of circulating hormones that facilitate pro-carcinogenic events[16].

Increased alcohol consumption has been identified as a risk factor for multiple different cancers, including breast cancer. It is particularly associated with increased risk of ER+ disease[17]. It has been suggested that this is because alcohol consumption increases levels of estrogen, leading to a hormonal imbalance that increases risk of carcinogenesis within the female organs[18]. Other explanations include a direct and indirect carcinogenic effect of alcohol metabolites and related impaired nutrient intake[19].

Smoking (both active and passive) also increases breast cancer risk, due to the carcinogens found in tobacco[20, 21]. Transported to the breast tissue, it has been suggested that these carcinogens increase the frequency of mutations within oncogenes and tumour suppressor genes (particularly, *TP53*) that, in turn, predispose to breast cancer development.

Other lifestyle factors are thought to play an effect as well, however, current data are insufficient to compare the results and draw credible data. Some studies attempt to identify whether certain vitamins might exhibit protective properties. This especially relates to vitamin D, since high serum levels of 25-hydroxyvitamin D are thought to be linked to a decreased breast cancer incidence rate[22], while intensified expression of vitamin D receptors was demonstrated to be related to lower mortality[23]. Other potential risk factors include chronic exposure to chemicals, intake of certain drugs and a diet rich in ultra-processed food. A 10% increase of ultra-processed food in the diet was found to be associated with an 11% greater risk of breast cancer[24]. However, data available for these factors remain inconsistent, and further evaluation is required to confirm their relationship with breast cancer risk.

1.1.2.2. Non-modifiable risk factors

Female sex

Female sex is the major factor associated with an increased risk of breast cancer, with over 99% of cases occurring in women. However, not all women are at equal risk, with the highest incidence rates observed among white non-Hispanic women[25]. The mortality rate, however, has been found to be significantly higher in black women who have also been reported to be more susceptible to the most aggressive breast cancer subtype, TNBC[26].

Older age

It is also well-established that older age is associated with higher risk of developing this disease. Around 80% of breast cancer cases occur in females aged over 50; and the 10-year probability of developing invasive breast cancer rises from < 1.5% at age 40, to around 3% at age 50 and > 4% by age 70, resulting in a cumulative lifetime risk of 13.2% (or 1 in 8)[27]. Interestingly, there is also a relationship between age and molecular subtype of breast cancer. For example, luminal A cancer is most frequently diagnosed in women over the age of 70, while TNBC is the most prevalent amongst patients < 40 years old[28].

Family history

Another important factor is family history – around 13-19% of breast cancer patients have a first-degree relative affected by the disease[29]. Importantly, the incidence risk is higher in all the patients with family history regardless of their age.

Reproductive factors

Many key female reproductive milestones have a strong association with the risk of breast cancer development. Multiple studies have reported a strong relationship between higher disease risk and exposure to endogenous hormones (particularly, estrogen and progesterone). Thus, events such as menarche, pregnancy, breastfeeding and menopause can significantly alter a woman's risk of breast cancer. For example, an early full-term pregnancy, especially < 20 years old, was associated with a 50% reduction in the risk of breast cancer development compared to nulliparous women[30, 31]. Subsequent pregnancies have been reported to carry lower protective effects, however, independent of maternal age. Interestingly, despite the overall protective effects of pregnancy, there is a short-term increase in risk of breast cancer immediately after parturition. Additionally, postpartum breast cancer patients have been shown to have higher risk of metastasis and worse clinical outcomes[32]. Longer duration of breastfeeding has also been associated with decreased risk of both receptor-positive and -negative tumours[33, 34]. Early age at menarche and late menopause are both associated with an increased breast cancer risk[35]. Additionally, hormone replacement therapy (HRT) often prescribed to relieve unpleasant menopause symptoms has been associated with an increased risk, especially when taken longer than 5-7 years[36]. This increased risk is, however, lost after treatment has been stopped for a period of 5 years or longer.

Density of breast tissue

Another strong risk factor is the density of breast tissue, known as mammographic density (MD) or percent density (PD)[37, 38]. PD – is the percentage of breast area appearing radiodense on a mammogram. PD is a composite of two phenotypes: the dense area (DA) and the nondense area (NDA). DA reflects the amount of fibroglandular tissue that attenuates X-rays more than fat and hence appears light (dense) on a mammogram. NDA, in turn, consists of predominantly fatty tissues and appears radiotranslucent or dark on a mammogram. Higher PD is observed in younger females, those with lower BMI as well as during pregnancy and breastfeeding. Generally, higher PD is strongly associated with increased breast cancer risk. Women with ≥ 75% density have a 4 to 5-fold greater risk compared to those with little or no dense tissue, independent of other known risk factors[39, 40]. Additionally, recent studies have demonstrated that NDA is associated with decreased disease risk independently of DA, suggesting that breast adipose tissues may have an

important role in normal mammary gland growth and function[41, 42]. However, the mechanisms underlying these associations remain poorly understood.

Genetics

There are eight Hallmarks of Cancer that comprise a set of functional capabilities acquired by human cells that are necessary for the development of malignant tumours[43]. These include sustaining proliferative signalling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing/accessing vasculature, activating invasion and metastasis, reprogramming cellular metabolism, and avoiding immune destruction. The acquisition of these hallmarks is made possible by two enabling characteristics: genome instability and tumour-promoting inflammation. Genome instability generates random (somatic) mutations; the acquisition of these mutations can be affected by pre-existing (germline) variants. Germline variants occur in gametes and can be passed onto offspring at the time of conception, while somatic variants occur in an individual cell during a person's lifetime and cannot be inherited. This work focuses on germline variants.

The first breast cancer risk-associated germline variants were identified in the *BRCA1* and *BRCA2* genes[44, 45]. Although the prevalence of truncating mutations in these genes is low in the general population (approximately 1 in 400), it has been estimated that the truncating mutations in these two genes account for ~ 17% of the familial relative risk (FRR) for early-onset breast cancer[46, 47]. Other genes for which highly penetrant mutations have been reported to be associated with breast cancer include *PTEN*, *TP53*, *CDH1* and *STK11*[48-51]. Additionally, variants in several DNA repair genes that interact with BRCA genes have been associated with breast cancer risk, however, these are characterised by a lower penetrance (moderate risk). These genes include *CHEK2*, *ATM*, *BRIP1* and *PALB2*[52-55]. Although many high- and moderate-penetrance genes have been identified, a substantial proportion of FRR is yet to be explained.

1.1.3. Key research areas

There are three main areas in breast cancer research – prevention, detection and treatment. Prevention research focuses on studying factors that affect an individual's predisposition to breast cancer, i.e., events that occur before the cancer develops. Detection research aims to develop and optimise biological tests and imaging techniques that will allow

earlier detection of cancers while minimising the invasiveness. Treatment research focuses on the development and improvement of treatment methods.

1.2. Genetic Variation

Completed in 2003, the Human Genome Project (HGP) was a 13-year international effort to decipher and publish the human genome[56]. The first 'full' version of the human reference genome covered 99% of the euchromatic genome, leaving important heterochromatic regions unfinished. Since then, the reference genome has continued to be updated and refined by the Genome Reference Consortium (GRC), however, until recently around 8% of the genome remained incomplete due to technological limitations. In 2022, the Telomere-to-Telomere (T2T) Consortium published a collection of papers that report the first truly complete 3.055 billion–base pair sequence of a human genome, that has gapless assemblies for all chromosomes except Y[57].

The reference genome is not that of an individual person, but instead intends to be a representation of the 'average' DNA sequence. It has been estimated that any individual's genome differs from the reference genome at 4.1 to 5 million sites[58]. Such germline genetic variation accounts for a substantial amount of phenotypic variation, including cancer predisposition.

The 1000 Genomes Project was launched in 2008 with the aim to identify all common human genetic variation. Since then, a series of papers have been published by the consortium, revealing the extent of genetic variation between individuals and populations. There are two main types of variants: single nucleotide variants (SNVs) and structural variants (SVs).

1.2.1. Single nucleotide variants

An SNV is a DNA sequence variation that occurs at a single nucleotide. If at least two alleles of the variation have frequencies of more than 1% in a large population of unrelated individuals, then the SNV is classified as a single nucleotide polymorphism (SNP)[59]. SNPs are estimated to occur at 1 out of every 300 base pairs, making them the most common type of human genetic variation. Depending on where a SNP occurs, it might have different consequences at the phenotypic level.

Less than 1% of SNPs occur in coding regions. Coding SNPs are categorised into two main groups: synonymous and non-synonymous. Synonymous SNPs lead to redundant changes in a codon, therefore, not affecting the coding sequence. Non-synonymous SNPs, in turn, lead to non-redundant codon changes, potentially resulting in an amino acid change (missense SNPs) or the introduction of a premature stop codon (nonsense SNP). Additionally, a single base deletion can result in a frameshift effect. Although changes to the primary protein structure can be predicted, functional effects of these SNPs are hard to predict from their sequence alone and, therefore, require targeted investigation.

The remainder of SNPs are found in non-protein-coding regions. Although initially non-coding DNA was considered to be largely 'junk', it is now becoming clear that much of it is integral to the function of cells, particularly the control of gene expression. For example, non-coding DNA contains sequences that act as regulatory elements, determining when and where genes are turned on and off. Despite its functionality, large parts of the non-coding genome remain to be characterised, making it difficult to predict and distinguish SNPs with a functional effect from those that are functionally silent. Given the regulatory nature of some non-coding regions (promoters, enhancer, silencers), it has been proposed that non-coding SNPs can disrupt such regulatory elements and modulate levels of gene expression. However, some non-coding sequences, such as miRNAs and lncRNAs, could be considered functional on their own, so more studies are required to investigate other potential mechanisms.

1.2.2. Structural variants

SVs are defined as the variants affecting regions of more than 1 kb. There are different types of SVs. Many are large insertions and deletions (indels), while others are inversions or more complex rearrangements. Some structural variants are copy number variants (CNVs). These occur when a region of the genome is duplicated (sometimes more than once). CNVs can vary in size, from microsatellite regions composed of long tranches of bi- or tri-nucleotide repeats to large regions encompassing genes. Although more than 99% of variants consist of SNPs and short indels, SVs affect more bases: the typical genome contains around 2,100 – 2,500 structural variants, affecting ~20 million bases of sequence[58].

SVs tend to occur in repetitive regions of the genome and show greater internal complexity, making them difficult to study. Therefore, the effects of SVs at a genomic level remain to be characterised. Some research, however, suggests that, despite the relatively large size, some SVs can be generally well-tolerated. For instance, Sudmant and colleagues identified 240 genes that were homozygously deleted in normal individuals without clear phenotypic effects[60].

1.3. Genome-wide Association Studies

Individual variants tend to account for a relatively small proportion of risk association. This can be explained by considering the variants in a context of evolution. Most variants arising with a particularly deleterious effect will typically undergo strong negative selection and removal from the gene pool, although, arguably, this may be less relevant to late onset diseases such as breast cancer.

The relatively subtle effects of common variants means that they are much more likely to be distributed throughout a population, rather than conspicuously inherited in families. Therefore, to investigate the role of variants in any phenotype of interest, large studies with high statistical power are needed.

Genome-wide association studies (GWAS) involve testing genetic variants across the genomes of many individuals to identify genotype-phenotype associations. Over the past decade, GWAS have revolutionised the field of complex disease genetics[61]. Since 2005, when the first GWAS for age-related macular degeneration was published[62], over 50,000 significant associations have been reported between genetic variants and common diseases and traits[63]. These associations have aided in the identification of novel disease-causing genes, biomarkers and drug targets.

GWAS start with identification of the disease (or trait) to be studied and selection of the appropriate study population (cases and controls for the disease, or an unselected population sample for the trait). Genotyping can be performed using microarrays, where the identities of many variants can be tested simultaneously and then combined with imputation in order to increase the density of markers. Association tests are used to select regions of the genome associated with the phenotype of interest, and meta-analysis is commonly performed to increase the statistical power to detect associations.

Due to the large number of variants in the genome, it is often neither feasible nor cost-effective to genotype each of them individually. Luckily, it is possible to impute a large number of variants to a high level of confidence by genotyping a much smaller number of 'index SNPs' due to linkage disequilibrium (LD).

1.3.1. Linkage disequilibrium

LD is the correlation between the neighbouring genetic variants in a population such that the allelic combinations of variants in LD are co-inherited more often than would be expected by chance if they were independent. Most commonly, LD arises through the process of sexual recombination. During meiosis, homologous chromosomes undergo a reciprocal exchange of DNA to allow for the variation in germ-cell lineages. Some regions are more likely to come together and to be passed to gametes as a unit. Therefore, variants that are in close proximity with each other are more likely to be co-inherited, with LD decreasing exponentially as distance increases.

There are several LD measures. Although selection of the most appropriate measure depends on the objective of the study, the two most widely used measures are r^2 and D'. Both these measures estimate the difference between the observed and expected gametic haplotype frequencies. Two fully correlated variants would have r^2 and D' scores of 1, while two non-correlated variants would have scores of 0. The main difference between the two measures is that r^2 (but not D') is influenced by allele frequencies. As a result, D' provides an indication of LD between two variants, while r^2 also gives an idea of how informative this association is for imputation based on relative frequencies of the variants.

Because of LD, it is possible to genotype a subset of variants (referred to as index SNPs) that capture a large proportion of local variation. Index SNPs are essentially proxies for all their correlated SNPs. This is useful for identifying risk loci at a genomic level but makes it difficult to discern the causal variant(s) – since index SNPs are only chosen based on their ability to capture regional genomic variation, there is no reason why they would be more likely to be functional than any other of their associated variants. Additionally, most association signals map to non-protein-coding regions of the genome, for which biological interpretation is challenging. Consequently, once a GWAS has been performed, additional steps are required to identify the causal variants and their target genes. In cases where a small number of variants are associated with risk, it may be

possible to perform functional investigation of all variants of interest. However, this approach becomes less feasible when larger numbers of variants are involved. As a result, prioritisation and shortlisting of credible causal variants (CCVs) are required before proceeding with in-depth functional characterisation.

1.3.2. Breast cancer GWAS

1.3.2.1. The Breast Cancer Association Consortium

The Breast Cancer Association Consortium (BCAC) is a forum of investigators interested in the inherited predisposition to breast cancer. Since its formation in 2005, the consortium has been responsible for a large number of both genome-wide and small-scale studies (https://bcac.ccge.medschl.cam.ac.uk/publications/). There are currently over 100 research groups that participate in BCAC.

1.3.2.2. COGS project

Later, BCAC combined with three other consortia – Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA), Ovarian Cancer Association Consortium (OCAC) and Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome (PRACTICAL) – to form the Collaborative Oncological Gene-environment Study (COGS) 'super-consortium'. The project was designed to improve understanding of genetic susceptibility to three hormone-related cancers: breast, ovarian and prostate cancers. The major strategy included replication of GWAS-identified associations, with secondary studies being focused on dense genotyping of SNPs for the fine mapping of associated regions.

The consortium worked together with Illumina to design a high-density, custom iSelect SNP genotyping array (called the iCOGS array) that would allow genotyping of the three cancers in large case-control studies[64]. The array is an Illumina Custom Infinium array which includes over 200,000 SNPs. The array allowed identification of many breast, ovarian and prostate cancer susceptibility regions, some of which overlapped suggesting shared mechanisms. The demonstrated benefits of the iCOGS array also informed the design of a next-generation cancer genotyping platform (called OncoArray) to identify risk variants for the five most common cancers.

1.3.2.3. OncoArray Consortium

The OncoArray Consortium brought together three consortia: BCAC, CIMBA and Genetic Associations and Mechanisms in Oncology Initiative Consortium ((GAME-ON), itself consisting of: Follow-up of Ovarian Cancer Genetic Association and Interaction Studies (FOCI); Colorectal Transdisciplinary Study (CORECT); Transdisciplinary Research in Cancer of the Lung (TRICL); Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) and Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE)). The overall goal of the OncoArray Consortium was to gain new insights into the genetic architecture and mechanisms underlying five common cancers: breast, ovarian, colorectal, lung and prostate cancers.

The Consortium designed a next-generation cancer genotyping microarray comprising 230,000 SNPs and used it to genotype 447,705 samples[65].

1.3.2.4. Breast cancer risk loci

Multiple breast cancer GWAS have been published over the last decade. Collectively, these studies have identified genetic variants associated with breast cancer risk in over 150 genomic regions, with the two most recent studies published in 2017 and 2020.

In the first of these, Michailidou and colleagues[66] genotyped 122,977 cases and 105,974 controls of European ancestry. They identified 65 novel loci in addition to confirming 77 previously discovered regions. To define a set of credible causal variants at the new loci, the authors selected all variants with *p* values within two orders of magnitude of the most significant SNPs in each region. Across the 65 novel regions, this identified 2,221 CCVs, while the 77 previously identified loci contained 2,232 CCVs.

In the second analysis, Fachal and colleagues[67] performed large-scale genetic fine-mapping of 150 breast cancer susceptibility regions in over 217,000 breast cancer cases and controls of European ancestry. Stepwise multinomial logistic regression was used to identify the number of independent risk signals within each region (Figure 1.1) and to define a set of CCVs for each signal (defined as variants with *p* values within two orders of magnitude of the index SNP at each signal). This resulted in the selection of 7,394 CCVs at 196 'strong-evidence' signals (defined as having association *p* values $< 10^{-6}$ after adjusting for other variants) across 129 genomic regions.

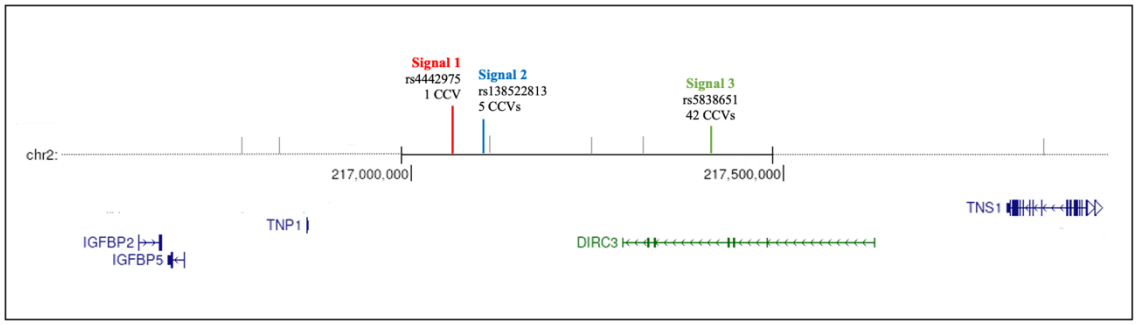

Figure 1.1: An example of a GWAS region having multiple independent risk signals. An overview of a breast cancer risk locus at 2q35. Fine-scale mapping of the 2q35 locus has defined three independent 'strong-evidence' signals (conditional $p < 1 \times 10^{-6}$) annotated by rs4442975 (signal 1; 1 CCV), rs138522813 (signal 2; 5 CCVs) and rs5838651 (signal 3; 42 CCVs). In addition, 6 'moderate' signals (grey) were identified at this region ($10^{-6} <$ conditional $p < 10^{-4}$). All coordinates are based on GRCh38/hg38.

Despite the identification of a large number of breast cancer risk loci, only a small proportion have been studied in detail. Follow up fine-mapping and functional studies have so far investigated only around 20 regions (Table 1.1).

Risk locus	Reference
1p11.2	Figueroa et al., 2011; Horne et al., 2016
2q33.1	Lin et al., 2015
2q35	Ghoussaini et al., 2014; Dryden et al., 2014; Wyszynski et al., 2016; Baxter et al., 2021
4q24	Guo et al., 2015
5p12	Milne et al., 2011; Ghoussaini et al., 2016
5p15.33	Bojesen et al., 2013
5q11.2	Glubb et al., 2015
6q25	Dunning et al., 2016
7q21	Milne et al., 2011
8p12	Glubb et al., 2020
8q24	Shi et al., 2016
9q31.2	Orr et al., 2015
10q26	Meyer et al., 2013
10q21.2	Darabi et al., 2015
11q13	Lambrechts et al., 2012; French et al., 2013; Betts et al., 2017
12p11	Zeng et al., 2016
12q24	Beesley et al., 2020
14q24.1	Figueroa et al., 2011; Lee et al., 2012
16q12	Udler et al., 2010
17q22	Darabi et al., 2016
19p13.1	Stevens et al., 2012; Lawrenson et al., 2016

Table 1.1: Functionally investigated breast cancer risk loci.

1.3.2.5. GWAS in diverse populations

To achieve effective imputation and association analysis, GWAS assume that all studied individuals have a similar LD structure. Consequently, most of the breast cancer GWAS performed to date have focused on European populations, mainly because they can be assessed in larger numbers. A few smaller studies have looked into East Asian and African populations[68-72]. Some of the signals identified in these studies are 'shared' signals, but they have also detected novel, population-specific signals.

Additional studies in diverse populations will be useful to identify more population-specific signals and to deepen our understanding of breast cancer risk. Different populations may also be predisposed to different breast cancer subtypes, so further studies may also aid in identification of subtype-specific signals.

1.3.2.6. Other relevant GWAS

Some studies have also investigated breast cancer risk related phenotypes, such as mammographic density[73-75] or age at menarche and menopause[76]. In addition, one study has looked specifically into associations with early-onset breast cancer[77].

1.4. Chromosome Conformation Capture

As discussed, breast cancer GWAS have identified approximately 200 'strong-evidence' independent risk signals[67]. Most of these signals map to non-protein-coding regions and are thought to affect transcriptional regulation[78-80], however, the causal variants and target genes mediating these associations remain largely unknown, with only a few regions studied in detail (Table 1.1). Attempts to understand the mechanisms by which these signals influence risk have been hampered by strong local correlation between multiple genetic variants which makes it difficult to distinguish causal variants from a large numbers of correlated variants.

Developed over a decade ago, Chromosome Conformation Capture (3C) technology allows the mapping of regulatory regions and identification of their respective target genes[81]. 3C-based technology detects the relative interaction frequency between two regions within the genome, from which chromatin folding can be inferred. Since its establishment, 3C has been further modified to increase throughput, leading to the

development of 4C[82, 83], 5C[84] and Hi-C[85]. Although 3C is a powerful technique for detecting physical interactions between regulatory elements and their targets, its limitation is that only interactions with pre-specified target fragments will be identified ('one-by-one' approach). 4C considers all interactions, however, only for a single region of interest (the bait fragment; 'one-by-all' approach) and 5C allows the analysis of multiple regions and their targets, but with a condition that both bait and targets are within a pre-specified region ('many-by-many').

1.4.1. Hi-C

A genome-wide version of 3C, called Hi-C, was introduced in 2009 by coupling 3C with massively parallel sequencing. One of the strongest advantages of Hi-C is its agnosticism. Unlike older chromosome conformation techniques, Hi-C requires no prior assumptions about interaction partners, with all interactions detected by sequencing ('all-by-all'). Today, Hi-C is the most commonly used 3C variant, that has been proven to be a useful tool not only for the identification of 3D genome folding patterns, but also for the *de novo* whole genome sequence assembly[86, 87] and translocation detection[88].

1.4.1.1. Description of the method

The various 3C-based techniques have four common steps: (1) formaldehyde crosslinking of chromosomes to covalently link spatially proximal chromatin segments; (2) fragmentation of DNA into smaller pieces; (3) ligation of linked DNA fragments under diluted conditions where intra-molecular ligation is strongly favoured over inter-molecular; (4) detection and quantification of ligation products. The main difference between 3C-based methods comes mostly from the last step. While in the 3C protocol, ligation products are identified one at a time using PCR with locus specific primers, Hi-C products are detected using next-generation sequencing. This became possible due to the incorporation of biotinylated nucleotides at the digested DNA ends prior to re-ligation, thereby allowing specific capture of chimeric molecules using streptavidin-coated beads.

Although a number of experimental parameters can vary, a typical Hi-C protocol is outlined below[89].

Cell crosslinking

The first step of any 3C-based method involves formaldehyde crosslinking of chromosomes to covalently link spatially proximal chromatin segments, specifically, DNA-DNA interactions bridged by proteins. Starting with a large number of cells is recommended in order to fully capture individual interactions (including the infrequent ones), resulting in complex, high-resolution Hi-C libraries. Although formaldehyde-based crosslinking biases have been proposed[90], it remains the 'gold-standard' in chromatin immunoprecipitation (ChIP) and 3C. Additionally, most of these biases can be removed using several normalisation techniques.

Cell lysis and chromatin digestion

Crosslinked cells are lysed in cold hypotonic buffer supplemented with protease inhibitors to maintain protein-DNA complexes. Next, lysed cells are incubated in Sodium Dodecyl Sulfate (SDS) to eliminate proteins that have not been crosslinked to DNA and to open the chromatin for a more efficient and homogeneous digestion.

Chromatin is then digested with a method of choice. The average fragment size is an important factor affecting the resolution of future libraries, so the choice of fragmentation method is important and depends on several parameters, such as the goal of the experiment or the region selected for the analysis (if applicable). Important factors to consider include desired resolution, spacing of the digestion sites and the overall digestion efficiency. Early protocols recommended digestion with 6-cutter enzymes, such as HindIII[91, 92]. HindIII digests the human genome to an average length of 3 – 4 kb, limiting the final library resolution to ~ 10 kb. Later, the use of 4-cutter enzymes, such as MboI and DpnII, has been proposed[89, 93]. These enzymes fragment DNA to an average length of ~ 500 bp which, in theory, could increase the resolution to ~ 1 kb. However, since the distribution of RE sites is uneven across the genome, some regions of interest may remain insufficiently covered. Therefore, to account for uneven and non-random digestion, methods were developed that use multiple restriction enzymes in combination (Arima Hi-C kit) or sequence-independent (RE-free) approaches. Two such examples are DNase I Hi-C[94, 95] and Micro-C[96]. It has been proposed that the DNase I Hi-C method, that uses DNase I for chromatin fragmentation, may be a more suitable option for probing interactions between regulatory elements, given that DNase I preferentially cleaves nucleosome-depleted regions. Micro-C, in turn, is suggested to be a complementary

approach, that is more suitable for assessing shorter-range interactions (between 200 bp and 4 kb) but at higher resolution.

Marking DNA ends with biotin

RE-based chromatin fragmentation generates an overhang that is subsequently filled in with deoxyribonucleotides. Replacing one deoxyribonucleotide (usually dATP or dCTP) with a biotin-conjugated variant marks the sites of digestion with biotin and allows further enrichment of these sites in a Hi-C library. It is this specific fill-in step that separates Hi-C from other 3C-based methods.

In the RE-free protocols (such as DNase Hi-C[94]), the biotin marking is performed by ligation of biotinylated-bridge adapters through T-A ligation. Because DNase I digestion produces a heterogeneous mixture of fragment ends composed of 5′- and 3′-overhangs of varying lengths as well as blunt ends, these ends have to be enzymatically repaired and dA-tailed prior to ligation of the bridge adapters.

In both cases the fill-in step is performed at low temperature, which is crucial for efficient incorporation of the large biotin-conjugated deoxyribonucleotides (or biotinylated-bridge adapters).

Ligation, reversal of crosslinking and DNA purification

The fourth step involves chromatin ligation. While older protocols used SDS to inactive the RE prior to ligation[91, 92], later it was replaced by heat inactivation, with the digestion, biotinylation and ligation being performed ‘in situ’ (i.e. within a permeabilized nucleus that is not lysed into the solution)[89, 93]. Avoiding high concentrations of SDS, nuclei lysis, and dilution to large volumes during digestion and ligation increases Hi-C reproducibility and quality as well as decreasing the capture of background interactions[97]. The generally accepted explanation for this has been that intact nuclei constrain the movement and random collisions of crosslinked complexes, but other factors could have an effect as well. For example, it has also been suggested that high concentrations of SDS and its subsequent Triton sequestration can result in aggregates of material that reduce digestion, fill-in and ligation efficiencies[98].

Ligation is performed at low DNA concentrations to strongly favour intra-molecular ligation of crosslinked fragments over background inter-molecular ligation between non-crosslinked fragments[99]. Because intra-molecular ligation is kinetically fast, ligation time should be kept to a minimum to avoid increasing background ligation or generation of circularised ligation products of single restriction fragments – these are not considered valid pairs and should be removed computationally.

When interacting fragments are ligated into chimeric pieces of DNA (di-tags), proteins that hold them in close proximity can be removed. This is achieved by thermal reversion of crosslinking in the presence of proteinase-K. After that, DNA is purified and prepared for sequencing.

Removal of biotin from un-ligated ends

In most Hi-C experiments a fraction of digested sites will have remained un-ligated[89]. These biotinylated but un-ligated ends (called dangling ends) may arise from incomplete fill-in of some overhangs (since in this case ligation to a proximal fragment will not occur) and the overall ligation will not be 100% efficient. These dangling ends are not informative and are not considered valid pairs. Some of them can be readily recognised (and removed) computationally, since both reads will map to a single restriction fragment. However, a sub-population of dangling ends can appear as valid interactions between adjacent restriction fragments. Dangling ends flanking an undigested restriction fragment (partial digest) will computationally be indistinguishable from a valid pair interaction with an inward orientation. Such read pairs increase in frequency with decreasing restriction fragment size (i.e., when more frequently cutting enzymes are used). Therefore, it is recommended to experimentally remove these dangling ends for two reasons. First, removal of dangling ends increases the proportion of informative intra-chromosomal reads, therefore helping to decrease the cost of sequencing by increasing the relative quantity of valid read pairs. Second, it eliminates dangling ends of partial digestion products that cannot be recognised and removed computationally.

This step is performed using T4 DNA polymerase and a low concentration of dNTPs to favour the 3’ to 5’ exonuclease activity over its 5’ to 3’ polymerase activity. By only providing dATP and dGTP, which are complementary to the inside of the overhang, the polymerase will not be able to complete re-filling the overhang after removing filled-in bases.

Although this step reduces the pulldown of a large fraction of unwanted fragments, some of them can remain insensitive to biotin removal. For example, if internally nicked DNA is repaired with biotinylated nucleotides during biotin incorporation, when too far away from DNA ends, these incorporated nucleotides will not be removed by T4 DNA polymerase.

Sonication, size-selection and end repair

For sequenced reads to be mapped correctly, each end of a read pair should not pass the chimeric ligation junction, since this will result in a sequence that cannot be mapped to a reference genome. Therefore, ligation products are sonicated to 200 – 650 bp (target size of 400 bp) in preparation for sequencing, since fragments of this length are likely to contain enough mappable sequence at each end before reaching a ligation junction.

Although sonication should result in a relatively small size range of fragments, an optional size-selection step can be performed to create an even tighter distribution of fragments. Size-selection is usually performed using Solid Phase Reversible Immobilization (SPRI) beads. SPRI beads are a mixture of magnetic beads and polyethylene glycol (PEG). SPRI beads decrease the solubility of DNA, because PEG (a crowding agent) occupies the hydrogen bonds of aqueous solutions. As a result of this crowding, DNA comes out of the solution and binds the beads. Larger molecules come out of the solution first, so the final concentration of PEG is used to generate a size cut-off.

DNA sonication causes damage of DNA ends, that have to be repaired with a mix of T4 and Klenow DNA polymerases, followed by a treatment with T4 polynucleotide kinase (PNK) that phosphorylates 5'-ends for subsequent A-tailing and adapter ligation.

Sequencing preparation

To enrich for Hi-C ligation junctions, streptavidin-coated beads with a high affinity for the incorporated biotin are used.

Next, Illumina paired-end (PE) sequencing adapters are ligated to both ends of the ligation products. Since the PE adapters are generated from DNA oligos, they have a 5'-dTTP overhang after duplexing, which increases ligation efficiency when presented with a free

3’ Adenyl. Ligation products are, therefore, adenylated using dATP and a Klenow fragment lacking 3’ to 5’ exonuclease activity, before adapters are ligated with T4 DNA ligase.

To get enough DNA for sequencing, the Hi-C library is amplified by PCR. However, it is important to avoid over-amplification of the library, since this will reduce its complexity.

Data analysis – defining valid read pairs

After sequencing, reads are mapped to a reference genome and valid interaction pairs are identified using read orientation.

Reads mapping to a single fragment, such as self-ligations, dangling ends or error pairs, are considered uninformative and can be identified by the read orientation (Figure 1.2). Outward pointing reads are classified as self-ligated fragments, inward pointing reads are considered dangling ends, and same-strand reads are defined as ‘error pairs’ (products that are a result of either a mis-mapping, random break, or an incorrect genome assembly products)[100].

Therefore, only reads that map to different fragments are used to assemble the Hi-C dataset. Although all four read strand combinations (inward, outward, same direction: left and right) are expected to be observed in equal proportions, in reality there is a bias towards inward read pairs. This bias is largely driven by very short-range interactions (genomic distance < 500 bp), which often represent dangling ends of partial digestion products – a type of invalid interactions that cannot be distinguished and filtered out computationally[89].

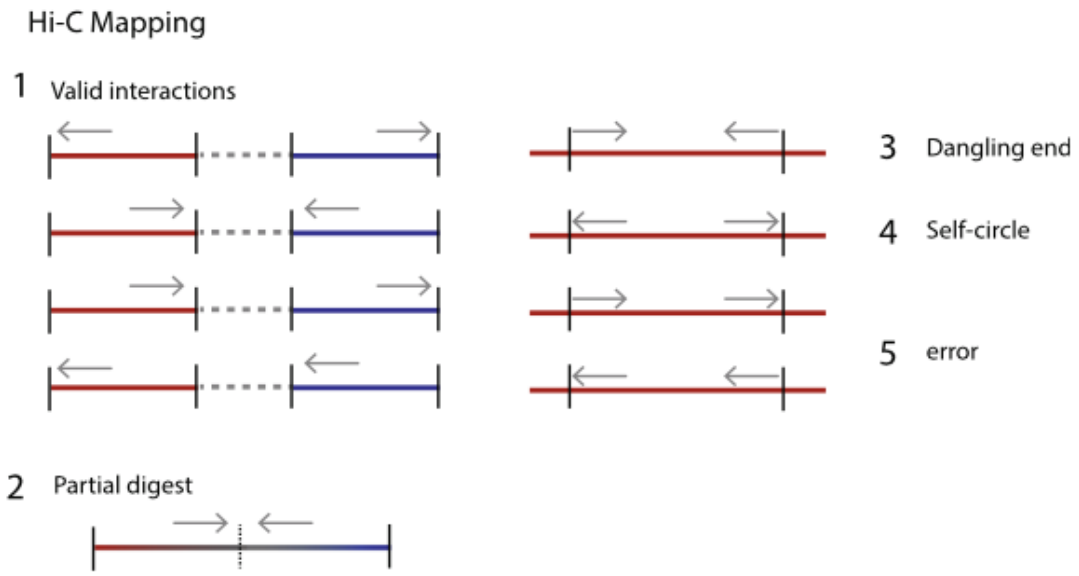

Figure 1.2: Possible Hi-C products. Following sequencing, the paired reads are mapped back to a reference genome and valid interaction pairs (1) are identified using read orientation. Although all four read orientations are possible and are expected to be observed in equal proportions, there is an imbalance towards inward read orientation, since some of these may be the result of undigested restriction sites (partial digest; 2). Only the reads that map to different fragments are used to assemble the Hi-C dataset, while the reads mapping to a single fragment (3 – 5) are considered uninformative. There are a few types of such uninformative reads: inward pointing reads are considered dangling ends (3), outward pointing reads represent self-circles (4), and same-strand reads are classified as 'error pairs' (5). Adapted from Belaghzal et al. (2017)[89].

1.4.1.2. Data resolution

Data resolution is one of the biggest challenges in Hi-C. Achieving sufficient coverage to support maximal resolution is difficult, because interaction space is very large. For instance, chromatin digestion using 6-cutter enzyme generates ~ 10^6 fragments in the human genome, resulting in an interaction space on the order of 10^{12} possible pairwise interactions[100].

The resolution of a Hi-C dataset depends on several factors, firstly – coverage. Higher sequencing depth allows coverage of more of the interaction space, thus improving the resolution. However, sequencing depth can be limited by a library complexity (defined as the total number of unique chimeric molecules that are present in a Hi-C library). Library complexity depends on a number of factors, including the number of cells used for library preparation, number of amplification cycles, etc. A low complexity library will saturate quickly with increasing sequencing depth, so less information will be gained from additional sequencing. The saturation curve of a library can be estimated by plotting the cumulative number of unique interactions observed versus increasing read depth.

Aggregation of restriction fragments into fixed-size bins allows the reduction of the interaction space, therefore increasing the resolution. Heatmaps generated using 100 – 500 kb-binned data allow the identification of large-scale genomic conformations (such as compartments). The location of topologically associating domains (TADs) can usually be identified using ~ 40 kb bins, while point-to-point interactions or loops can only be seen when data is binned at 10 kb or less[89]. When choosing a desirable bin size, it is important to consider the average fragment length in a library to reduce the number of bins containing no fragment ends. However, it is important to note that enzymes, such as HindIII and DpnII, result in a non-normal distribution of fragments in the human genome, so the average fragment length does not imply that most fragments are around this length. For example, when human genome is digested with HindIII (average fragment length 3 – 4 kb), nearly 30% of fragments are actually ≤ 1 kb, so some interactions can be detected at higher resolution than the average fragment length. Conversely, around 23% of fragments are > 5 kb, meaning that a bin size of 5 kb (to match the average fragment length) will result in many bins with no information.

1.4.2. Capture Hi-C

Initially, the resolution (1 – 10 Mb) prohibited the use of Hi-C for the interrogation of GWAS risk loci. To overcome this limitation, a version of Hi-C, called Capture Hi-C (CHi-C)[101], was developed (Figure 1.3). By incorporation of a target enrichment (capture) step, CHi-C allows the analysis of the subset of interactions for which the bait fragment maps to a pre-defined genomic region and the location of the target end is unrestricted ('many-by-all').

Although several target enrichment methodologies exist, in-solution hybridisation-based methods are often used due to their simplicity, efficiency, scalability and reproducibility[102]. In these methods RNA or DNA oligonucleotide baits (probes) are directed to the ends of targeted DNA fragments. The probes are biotinylated and can be recovered using streptavidin-coated magnetic beads to enrich for ligation events prior to next-generation sequencing.

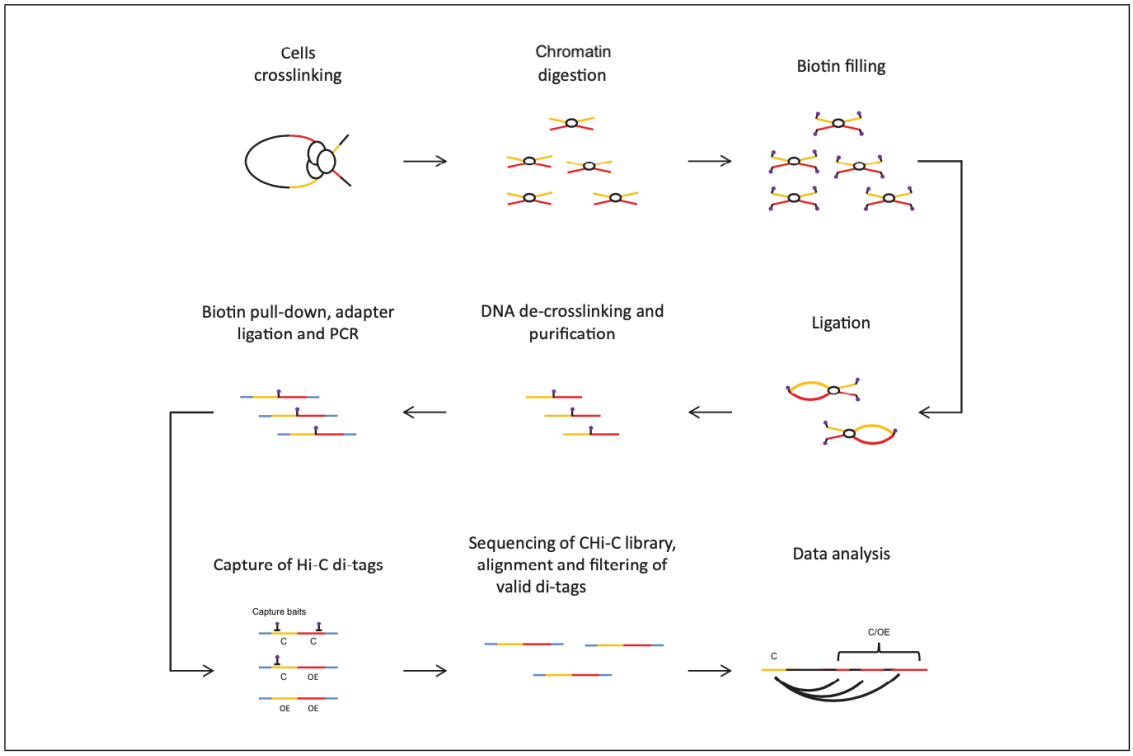

Figure 1.3: General Capture Hi-C workflow. C – captured, OE – other end. Adapted from Orlando et al. (2018)[103].

Although, in theory, capture baits can be designed to any regions of interest, two most common CHi-C types are promoter Capture Hi-C (pCHi-C) and region Capture Hi-C (rCHi-C). In pCHi-C, baits are designed to annotated promoters, resulting in strong enrichment of promoter-anchored interactions, while in rCHi-C, baits are designed to restriction fragments or LD blocks containing genetic variants associated with the disease of interest.

1.4.2.1. Protocols used in this project

CHi-C protocols are usually based on Hi-C protocols, which are extended to include a capture step. A number of different protocols and kits are available, however, in this project, I focused on three protocols: our standard in-house protocol, the Arima Genomics Hi-C kit and the Dovetail Genomics Omni-C kit.

Our standard in-house protocol (Standard, hereafter) is based on the Belaghzal and colleagues[89] and Orlando and colleagues[103] protocols and uses the HindIII restriction enzyme for DNA fragmentation. The average resolution of the HindIII digested libraries is ~10 kb. This protocol has been proven to generate informative CHi-C libraries;

however, it is time-consuming and requires large amounts of starting material, therefore complicating its usage for library generation from some types of primary cells.

The Arima Hi-C kit (Arima) is a highly simplified protocol, which requires significantly less time and starting material. The Arima RE mix comprises two enzymes: a 4-cutter enzyme that recognises the sequence GATC and a 5-cutter enzyme recognising GANTC. The use of two frequent cutter enzymes, especially with one that has a variable nucleotide within the recognition sequence should result in increased library resolution throughout the entire genome.

The Dovetail Genomics Omni-C kit (Dovetail) is quite similar to the Arima Hi-C kit in terms of time and starting material requirements. However, it benefits from a sequence-independent endonuclease approach to DNA fragmentation, which should lead to increased genomic coverage and reduced restriction enzyme density biases. Another advantage of this protocol is that it uses two crosslinking agents – formaldehyde and disuccinimidyl glutarate (DSG). Using formaldehyde in combination with DSG has been shown to better preserve chromatin contacts and increase data quality at various resolutions when compared to the use of formaldehyde alone[104].

1.4.2.2. Limitations of previously published breast cancer CHi-C studies

CHi-C data has been used for the generation of 'prioritised' lists of putative target genes and, to a lesser extent, CCVs, potentially providing an insight into which variants and genes influence breast cancer risk. Most breast cancer-related studies, however, have used HindIII digested libraries, which result in an average resolution of 10 kb. Additionally, due to the uneven distribution of digestion sites within the genome, some regions of interest have remained poorly covered. For example, at the 11p15.5 breast cancer risk locus all of the CCVs selected by the BCAC fine-mapping project[67] map to a large ~27 kb HindIII fragment, and, therefore, cannot be adequately resolved. Finally, due to the fact that breast cancer originates in epithelial cells[105], CHi-C data have only been generated in breast cancer and immortalised 'normal' breast epithelial cell lines. However, the behaviour of primary tumours is influenced by many different cell types as well as noncellular factors, in particular the tumour stroma has been shown to have a profound effect on cancer progression and may well influence tumour initiation too[106].

1.4.2.3. Project aims

Based on the information summarised above, my project aims were to:

1. Generate rCHi-C libraries in breast epithelial and fibroblast cell lines using three different protocols to identify and optimise the most suitable method for library generation in primary cells;
2. Generate higher resolution rCHi-C data in two types of primary breast cells (luminal epithelial cells and fibroblasts) to identify regulatory variants and target genes influencing breast cancer risk;
3. Compare cell line data to the primary cell data to evaluate the usefulness of cell lines as model systems;
4. Generate cell line pCHi-C data to validate rCHi-C findings and to identify 'indirect' interactions;
5. Using data generated in primary breast fibroblasts, determine whether a subset of breast cancer loci may act via the stroma to influence the risk.

2. Materials and Methods

2.1. Cell Culture

2.1.1. Cell lines

Breast cancer cell line T-47D (HTB-133) was obtained from the American Type Culture Collection (ATCC; LGC Standards). Reduction mammoplasty transformed normal breast fibroblast cell line GS2 was provided by Professor Clare Isacke (The Institute of Cancer Research, UK).

T-47D cells were cultured in Roswell Park Memorial Institute (RPMI) 1640 medium (Gibco, 11875093) supplemented with 10% foetal bovine serum (FBS; Gibco, 10500064), 10 µg/ml recombinant human insulin (Sigma, I9278) and 1% penicillin-streptomycin (Sigma, P4458) at 37°C, 5% CO_2.

GS2 cells were cultured in Dulbecco's Modified Eagle Medium (DMEM; Gibco, 11995065) supplemented with 10% FBS at 37°C, 10% CO_2.

Medium was changed twice a week. As cells approached confluency, they were washed with phosphate buffered saline (PBS; in-house) before addition of 0.25% trypsin-EDTA solution (Gibco, 25200056) to detach the cells. Trypsin was neutralised by addition of an excess of serum-containing growth media. Detached cells were pelleted by centrifugation at 200 x g for 5 minutes. After supernatant was discarded, cells were resuspended in fresh growth medium and seeded into new flasks.

Both cell lines were regularly tested for *Mycoplasma* contamination using an in-house service.

2.1.2. Primary cells

Primary breast luminal epithelial and fibroblast cells were provided by the Breast Cancer Now Tissue Bank (Dr. Jenny Gomm; Dr. Iain Goulding). These cells were isolated from two healthy premenopausal women undergoing reduction mammoplasty [patient IDs 1989N (age at surgery: 33 years) and 3002N (age at surgery: 31 years)].

Cells were cultured in DMEM/F12 medium (Sigma, D8437) supplemented with 10% FBS (Gibco, 10500056), 1% penicillin-streptomycin (Sigma, P4333) and 2.5 µg/ml amphotericin B (Sigma, A2942) at 37°C, 5% CO_2.

Medium was changed twice a week. As cells approached confluency, they were washed with PBS (Sigma, D8537) and detached by addition of 0.25% trypsin-EDTA solution (Cytiva, SV30031.01) diluted to 0.05% with PBS and incubation at 37°C. Trypsin was neutralised by addition of an excess of serum-containing growth media. Detached cells were pelleted by centrifugation at 380 x g for 3 minutes. After supernatant was discarded, cells were resuspended in fresh medium and seeded into new flasks.

Cells were regularly tested for *Mycoplasma* contamination using an in-house service.

2.2. Standard (in-house) Hi-C Protocol

Standard Hi-C library generation was performed as described by Belaghzal and colleagues[89] with some modifications. The adjusted protocol is detailed below.

2.2.1. Formaldehyde crosslinking

T-47D and GS2 cell lines were grown as described in Section 2.1.1. Since serum is very rich in proteins and therefore can affect the crosslinking efficiency by competing for formaldehyde, growth medium was replaced with Hanks' Balanced Salt Solution (HBSS; Gibco, 14175053) before fixation. Crosslinking of 20 million adherent cells was performed by adding 16% formaldehyde (Agar Scientific, R1026) to a final concentration of 1% and incubating the flasks for 10 minutes at room temperature (RT). Reaction was quenched by addition of glycine (VWR Chemicals, 101194M) to a final concentration of 128 mM. Cells were scraped off the culture flask and washed with PBS (in-house). Pelleted cells were snap-frozen in liquid nitrogen and stored at -80°C before continuing to the cell lysis.

2.2.2. Cell lysis and in-situ digestion

Cell lysis was performed in hypotonic lysis buffer (10 mM Tris-HCl pH8.0, 10 mM NaCl, 0.2% IGEPAL CA-630 (Sigma, I8896)) supplemented with 10µl 100X Halt Protease Inhibitor Cocktail (Thermo Scientific, 87786) to maintain Protein-DNA complexes.

Crosslinked cells were thawed, resuspended in 1 ml ice-cold buffer and incubated on ice for 15 minutes.

Cells were lysed on ice using a glass dounce homogeniser (two rounds; 30 strokes each). Permeabilised cells were pelleted at 2500 x g for 5 minutes and washed twice with 500 µl ice-cold 1X NEBuffer 2.1 (NEB, B7202S). After removing the second wash, cells were resuspended in 1030 µl 1X NEBuffer 2.1 and aliquoted into three tubes to final volumes of 342 µl. Each tube was used to prepare a separate Hi-C library replicate. The remaining 4 µl were used to check effective cell lysis using a disposable haemocytometer (Invitrogen, C10283).

Before digestion, lysed cells were incubated with 0.1% SDS (Ambion, AM9820) at 65°C for 10 minutes to eliminate non-crosslinked proteins and to open the chromatin for a better and more homogenous digestion. The reaction was terminated by addition of Triton X-100 (Sigma, X100-500ml) to a 1% final concentration.

To ensure maximal digestion, chromatin was incubated with 1500 U HindIII (NEB, R0104M) at 37°C overnight in a thermomixer with 950 rpm agitation. Digestion was terminated by heat inactivation of the restriction enzyme at 65°C for 20 min.

2.2.3. Marking of DNA ends with Biotin-14-dATP

5’ overhangs generated during DNA digestion were filled in with a Fill-in mix consisting of 1.5 µl each of 10 mM dCTP, 10 mM dGTP and 10 mM dTTP (Invitrogen, 10297018), 37.5 µl 0.4 mM biotin-14-dATP (Invitrogen, 19524016), 10 µl 5 U/µl DNA Polymerase I, Large (Klenow) Fragment (NEB, M0210S), 6 µl 10X NEBuffer 2.1 and 2 µl water. This reaction was incubated for 4 hours at 23°C in a thermomixer with interval agitation (900 rpm; 15 seconds every 5 minutes) before being returned to ice. This low temperature is crucial for efficient incorporation of the large biotinylated dATP and decreases 3’ → 5’ exonuclease activity.

2.2.4. Blunt end ligation

Ligation mix (120 µl 10% Triton X-100, 240 µl 5X ligation buffer (Invitrogen, 46300-018), 12 µl 10 mg/ml bovine serum albumin (BSA; NEB, B9000S), 50 µl T4 DNA ligase (Invitrogen, 15224090) and 243 µl water) was added to each tube, followed by incubation

for 4 hours at 16°C in a thermomixer with interval agitation (900 rpm; 15 seconds every 5 minutes).

2.2.5. Reverse crosslinking

Now that interacting loci are ligated into chimeric pieces of DNA, proteins that hold interacting fragments in close proximity can be removed. To do so, reactions were incubated overnight at 65°C with 50 µl 10 mg/ml proteinase K (Ambion, AM2546). After addition of another 50 µl 10 mg/ml proteinase K the following morning, reactions were incubated for a further 2 hours at the same temperature.

2.2.6. DNA purification

Reactions were incubated at 37°C for 15 minutes before being transferred to 50 ml falcon tubes. DNA was isolated using phenol:chloroform extraction followed by precipitation using a standard sodium acetate plus ethanol protocol.

Briefly, after addition of 1.3 ml phenol-chloroform-isoamyl alcohol (Invitrogen, 15593031), reactions were vortexed for 30 seconds, transferred to the Phase Lock Gel (Light) tubes (VWR Chemicals, 733-2477) and centrifuged according to the manufacturer's instructions (13000 rpm for 5 minutes). The top aqueous phase containing the DNA was transferred into a fresh tube, mixed with 0.1 reaction-volume of 3M sodium acetate, pH 5.2 (Sigma, S7899-100ml) and 2.5 reaction-volumes of ice-cold absolute ethanol and incubated for 3 hours at -20°C.

Precipitated DNA was pelleted by centrifugation at 3500 rpm for 30 minutes, washed with 70% ethanol and air-dried at 37°C for 15-30 minutes. The DNA pellet was resuspended in 100 µl water and incubated at 37°C overnight.

To degrade residual RNA, 1 µl 1 mg/ml RNase A (Invitrogen, AM2270) was added the following morning, and reactions were incubated at 37°C for additional 30 minutes. DNA was quantified using Invitrogen Qubit fluorometer and stored at -20°C.

2.2.7. Library quality check

To verify Hi-C junction marking with biotin-14-dATP and Hi-C ligation efficiency, a PCR digest assay can be performed (Figure 2.1). This involves PCR amplification of a particular ligation product formed by two distant restriction fragments, followed by its overnight digestion with HindIII, NheI or both. As distant fragments are targeted with the PCR primers, the product can only form when properly ligated chimeras are in place.

For the PCR amplification, AmpliTaq polymerase (Applied Biosystems, N8080161) was used according to the manufacturer's instructions. For a 50 µl standard reaction, the following were added to 250 ng of Hi-C DNA: 5 µl 10X PCR buffer, 4 µl 25 mM $MgCl_2$, 1 µl 10 mM dNTP mix, 0.5 µl 100 µM each 2F and 1R primers (Table 2.1), 0.25 µl 5 U/µl AmpliTaq DNA polymerase and water to a total volume of 50 µl. Thermocycle parameters are shown in Table 2.2.

Primer name	Sequence
Histone H1 HindIII Region 1 Reverse (1R)	GAAGAATAACAGCCGCATCAAAC
Histone H1 HindIII Region 2 Forward (2F)	GGCTGTGGTACCTGTAAAGAACTAACTC

Table 2.1: PCR digest assay primer sequences. Primers were designed by Laura Broome.

Step	Temperature	Time	Cycles
1	95°C	15 min	1
2	60°C	1 min	36
3	72°C	1 min	
4	94°C	30 sec	
5	60°C	2 min	1
6	72°C	10 min	1
7	4°C	∞	1

Table 2.2: PCR digest assay thermocycle conditions.

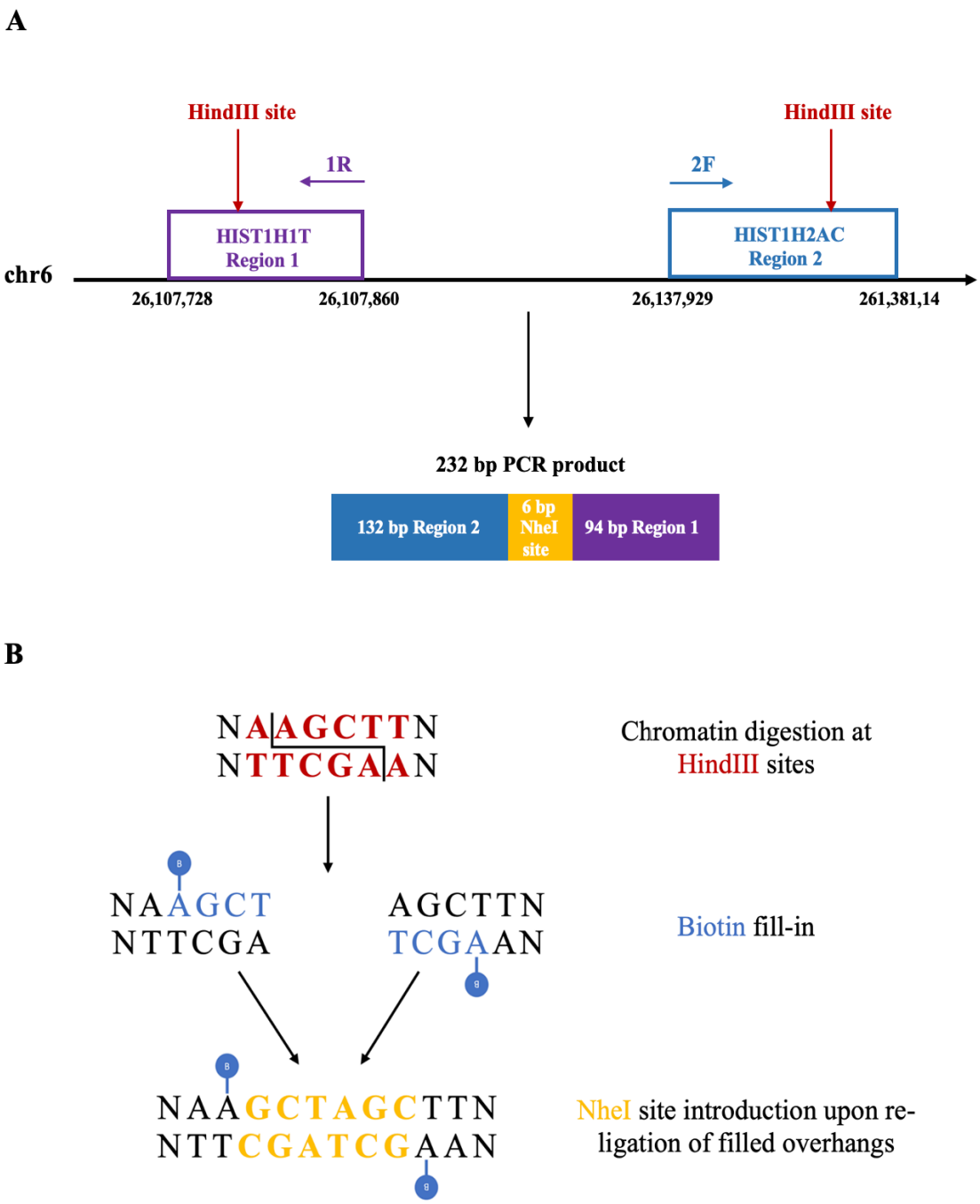

Figure 2.1: PCR digest assay for the quality control of standard Hi-C libraries. (A) In Hi-C libraries, a DNA long-range interaction occurs such that 1R and 2F primers (Table 2.1) can be used together to amplify a chimeric ligation product formed by two distant (~30 kb) fragments of *HIST1H2AC* Region 2 and *HIST1H1T* Region 1 with a 6 bp restriction enzyme site in between; (B) The biotin fill-in stage of Hi-C library preparation, followed by re-ligation, leads to the generation of a NheI restriction site in place of the original HindIII site, so only properly filled-in ligation products can be digested with NheI. Thus, comparison of the NheI digested and undigested fractions allows the estimation of the efficiency of these steps. Blue attachments represent biotinylation.

Each digestion reaction consisted of 15 µl PCR product, 2 µl buffer, 0.5 µl 20 U/µl HindIII (NEB, R0104S) or 0.5 µl 20 U/µl NheI (NEB, R0131S) (or both) and water to a total volume of 20 µl. Digestion was performed at 37°C for 2 hours, after which each reaction was run on a 2.5% agarose gel to estimate relative numbers of ligation events by quantifying cut and uncut bands (Figure 2.2).

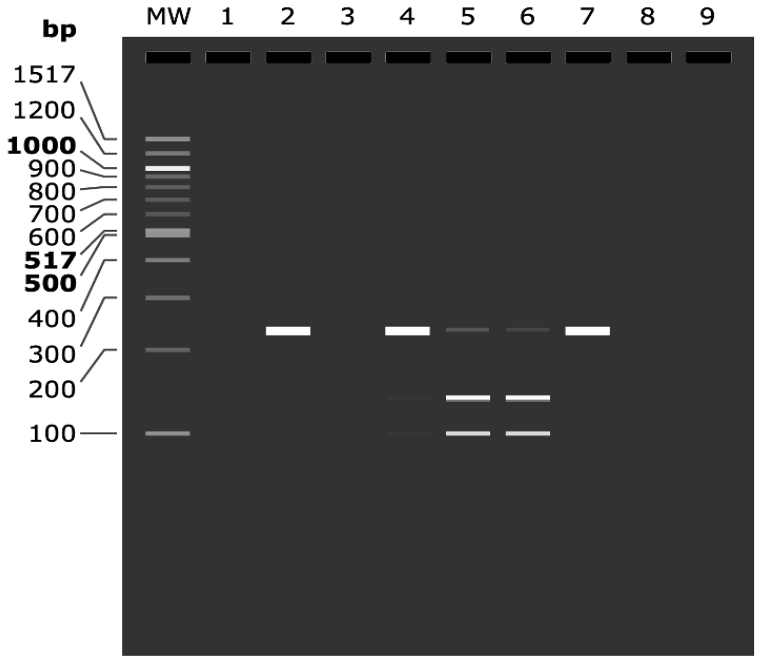

Figure 2.2: Quality control of Standard Hi-C libraries. An example of a typical PCR digest quality control gel, generated *in silico* using SnapGene. An *in silico* image was generated due to the low quality of images taken using the Syngene U:Genius[3] gel imaging system. (2) PCR product amplified with 2F and 1R primers (Table 2.1); PCR product digested with HindIII (4), NheI (5), both enzymes (6) or none (7, control). The molecular weight ladder used is the 100 bp DNA Ladder from NEB (MW).

2.2.8. Removal of biotin from un-ligated ends

In most Hi-C experiments some digested biotinylated sites will have remained unligated. To avoid pulling down such sites, biotin overhangs were removed using T4 DNA polymerase. A total of 80 µg Hi-C DNA was split into 16 reactions each consisting of 5 µg Hi-C library, 5 µl 10X NEBuffer 2.1, 0.125 µl 10 mM dATP, 0.125 µl 10 mM dGTP and 5 µl 3 U/µl T4 DNA polymerase (NEB, M0203L) in a total volume of 50 µl, and incubated at 20°C for 4 hours.

The biotin removal reaction was stopped by inactivating the enzyme at 75°C for 20 minutes. After cooling on ice to 4°C, 16 reactions were pooled into one and DNA was isolated using phenol:chloroform extraction followed by precipitation using a standard

sodium acetate plus ethanol protocol, as described in Section 2.2.6. The DNA pellet was resuspended in 750 µl 10 mM Tris-HCl, pH8.0 and incubated at RT for 1 hour before proceeding to DNA shearing.

2.2.9. DNA sonication and double-sided size-selection

DNA sonication and double-sided size-selection were performed as described by Orlando and colleagues[103].

Biotinylated DNA was split equally between six Covaris micro tubes (Covaris, 520052) and sheared to a target fragment size of 400 bp using Covaris LE220 machine with the following settings: fill level 6, duty cycle 10%, cycles/burst 200, peak incident power 175, time 60 seconds.

For the first step of size-selection, the total volume of each sonicated sample was brought to 200 µl. Next, 110 µl Ampure XP beads (Beckman Coulter, A63880) were added to each sample followed by a 20-minute incubation at RT. Tubes were then placed on a magnetic particle separator (MPS) and the unbound supernatant containing the DNA in the desired size range (< 700 to 1000 bp) was recovered.

In the second size-selection step, 240 µl AMPure XP beads were concentrated by placing the tube on the MPS and removing 180 µl of the supernatant. The beads were resuspended in the remaining 60 µl volume and added to the DNA recovered from the first size-selection step. After 10-minute incubation at RT, samples were placed on the MPS and supernatant containing fragments < 200 bp was discarded. The beads bound by DNA fragments between 200 and 650 bp were washed twice with 80% ethanol. The six aliquots were combined by resuspending all the beads in a total of 200 µl 10 mM Tris-HCl, pH8.0 and incubated at RT for 5 minutes, after which supernatant, containing the size-selected DNA, was collected from the MPS. 100 µl 10 mM Tris-HCl, pH8.0 were added to reach 300 µl total reaction volume required for the next step.

Sonication and size-selection efficiencies were checked by running the samples on the Agilent 2100 Bioanalyzer instrument using High-Sensitivity DNA kit (Agilent, 5067-4626) according to the manufacturer's instructions. A representative example of the expected Bioanalyzer profiles is shown in Figure 2.3.

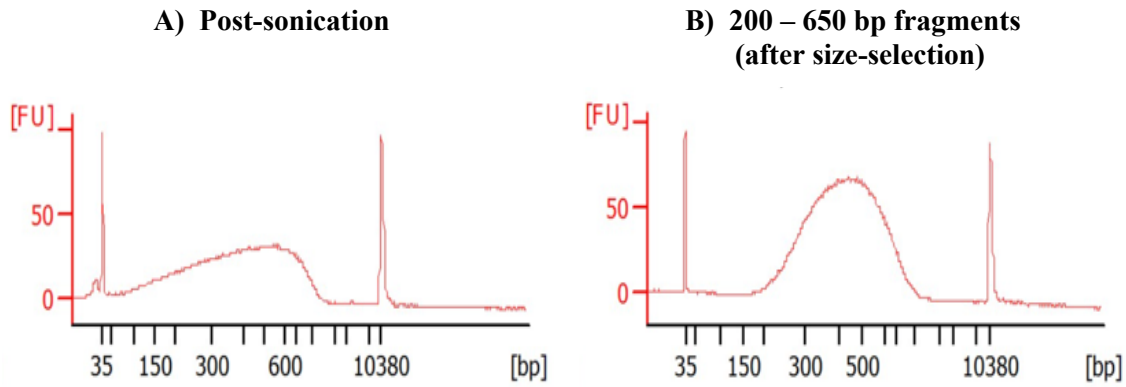

Figure 2.3: Post-sonication and double-sided size-selection Bioanalyzer profiles. Diluted T-47D DNA (1 in 10) was loaded on DNA High-Sensitivity chip and run on the 2100 Bioanalyzer instrument to show efficient sonication (A) and enrichment of the 200 to 650 bp fragments after successful size-selection (B).

2.2.10. Biotin pull-down

To enrich for Hi-C ligation junctions, streptavidin-coated beads with a high affinity for the incorporated biotin are used. To prepare the beads, 150 µl 10 mg/ml Dynabeads MyOne Streptavidin C1 beads (Life Technologies, 65001) were washed twice with 400 µl 1X Tween buffer (5 mM Tris-HCl, pH8.0, 0.5 mM EDTA, 1 M NaCl, 0.05% Tween). For each wash, beads were resuspended in fresh buffer, transferred to a new tube, rotated for 3 minutes at RT and then reclaimed against the MPS.

After removal of the second wash, beads were resuspended in 300 µl 2X No-Tween buffer (10 mM Tris-HCl, pH8.0, 1 mM EDTA, 2 M NaCl) and mixed with 300 µl library DNA. The mixture was incubated on a rotator at RT for 30 minutes, before reclaiming DNA-bound beads against the MPS. Next, beads were washed twice again in 1X Tween buffer. For each wash, beads were resuspended in 600 µl 1X Tween buffer, transferred to a new tube, incubated at 55°C for 2 minutes with shaking at 900 rpm and reclaimed against the MPS. For a final wash, beads were resuspended in 100 µl 1X T4 DNA ligase buffer (NEB, B0202S), transferred to a new tube and reclaimed against the MPS.

2.2.11. End repair, A-tailing and adapter ligation

To repair DNA ends damaged after sonication, beads were resuspended in a mix containing 88 µl 1X T4 DNA ligase buffer, 2 µl 25 mM dNTP mix, 4 µl 3 U/µl T4 DNA polymerase, 5 µl 10 U/µl T4 Polynucleotide kinase (NEB, M0201S) and 1 µl 5 U/µl DNA

Polymerase I, Large (Klenow) Fragment, and incubated at 20°C for 30 minutes. DNA-bound beads were reclaimed against the MPS and washed twice in 600 µl 1X Tween buffer (as described in Section 2.2.10). Next, beads were resuspended in 100 µl 1X NEBuffer 2.1 and reclaimed against the MPS.

To increase ligation efficiency of the Illumina paired-end adapters, the 3' ends of the ligation products need to be adenylated. To do so, beads were resuspended in a mix containing 90 µl 1X NEBuffer 2.1, 5 µl 10 mM dATP and 5 µl 5 U/µl Klenow Fragment (3'→5' exo-) (NEB, M0212S), incubated at 37°C for 30 minutes and reclaimed against the MPS. Then, beads were washed twice in 600 µl 1X Tween buffer and once in 100 µl 1X T4 DNA ligase buffer (as described in Section 2.2.10), before being resuspended in 50 µl 1X T4 DNA ligase buffer.

To prepare 15 µM annealed paired-end adapters required for the adapter ligation step, lyophilised top and bottom adapters (Table 2.3) were resuspended in T4 Polynucleotide Kinase buffer to a final concentration of 100 µM. 10 µl of each adapter were combined into a PCR tube and annealed as described in Table 2.4. To obtain a 15 µM working concentration, 46.7 µl 10 mM Tris-HCl, pH8.0 were added.

Adapter Name	Sequence
Top Adapter	ACACTCTTTCCCTACACGACGCTCTTCCGATC*T
Bottom Adapter	P-GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG

Table 2.3: Illumina paired-end adapter sequences. Top and bottom adapters have to be annealed together into a Y-shape structure before being ligated to both ends of the Hi-C library DNA. The bottom adapter is 5'-phosphorylated (P) in order to promote ligation. The top adapter has a phosphorothioate bond (*), resistant to exonuclease degradation.

Step	Temperature	Time
1	37°C	30 min
2	+ 0.5°C/sec to 97.5°C	
3	97.5°C	155 sec
4	- 0.1°C/5 sec to 20°C	
5	4°C	∞

Table 2.4: Illumina paired-end adapter annealing Thermocycle conditions.

Next, 3 µl of 15 µM annealed adapters were ligated to the Hi-C library DNA by adding 3 µl 400 U/µl T4 DNA ligase and incubating the reaction at 20°C for 2 hours.

Excess adapters were removed by reclaiming the beads against the MPS and washing them twice with 600 µl 1X Tween buffer (as described in Section 2.2.10) and once with 100 µl 10 mM Tris-HCl, pH8.0. Finally, beads were resuspended in 50 µl 10 mM Tris-HCl, pH8.0.

2.2.12. Pre-hybridisation PCR

To generate approximately 500 – 750 ng of DNA required for the target enrichment step, a Hi-C library must be partially amplified using adapter-specific primers (Table 2.5). To identify the optimal number of PCR cycles, test PCR reactions with 8, 10 and 12 cycles were carried out for each library (2 µl on-bead DNA, 0.4 µl 25 µM Prehyb Forward primer, 0.4 µl 25 µM Prehyb Reverse primer, 12.5 µl NEBNext High-Fidelity 2X PCR Mastermix (NEB, M0541S) and 9.7 µl water). PCR Thermocycle conditions can be found in Table 2.6.

Unbound supernatant containing amplified DNA was recovered using the MPS and purified with the MinElute PCR Purification kit (Qiagen, 28004). Test PCR product was quantified using the Agilent High-Sensitivity DNA kit with the Agilent 2100 Bioanalyzer system and optimal cycle number for each library was calculated.

Primer Name	Sequence
Prehyb Forward	ACACTCTTTCCCTACACGACGCTCTTCCGATC*T
Prehyb Reverse	CTCGGCATTCCTGCTGAACCGCTCTTCCGATC*T

Table 2.5: Pre-hybridisation PCR primers. * indicates phosphorothioate bond, resistant to exonuclease degradation.

Step	Temperature	Time	Cycles
1	98°C	30 sec	1
2	98°C	10 sec	8-12
3	65°C	30 sec	
4	72°C	30 sec	
5	72°C	5 min	1
6	4°C	∞	1

Table 2.6: Pre-hybridisation PCR Thermocycle conditions.

Next, multiple PCR reactions were set up for each library as described above, using the optimal cycle number. PCR product from these reactions was pooled and supernatant was recovered using the MPS. This supernatant was then cleaned using Agencourt AMPure

XP beads. For each clean up, 1.8X reaction-volume of beads was added to the tube followed by 5-minute incubation at RT. DNA-bound beads were then reclaimed against the MPS and washed twice with 800 µl 80% ethanol. Beads were dried at RT for 5 minutes before being resuspended in 80 µl nuclease-free water. After 5-minute incubation at RT, eluted DNA was recovered using the MPS and quantified using the Agilent High-Sensitivity DNA kit with the Agilent 2100 Bioanalyzer system before proceeding with the target capture (Section 2.6.1).

2.3. Arima Genomics Hi-C

To generate the Arima Hi-C libraries, the Arima Hi-C Kit (Arima Genomics, A410030) was used according to the manufacturer's instructions. Proximally ligated DNA samples were sheared as described in Section 2.2.9 before proceeding with the library preparation. Since the target capture step (performed following the Arima Hi-C library generation, and not part of the original Arima protocol) creates additional DNA loss, the size-selection step was omitted for these libraries to preserve DNA and to retain library complexity. Library amplification (pre-hybridisation PCR) was performed as described in Section 2.2.12. Different adapters were used in the Arima Hi-C kit; therefore, pre-hybridisation PCR primers were re-designed to be compatible with these libraries (Arima Fw and Arima Rv primers from Table 2.8). Arima Rv primer is complementary to the P7 sequence, while Arima Fw primer is complementary to P5 sequence and to the rest of the adapter sequence preceding the i5 index.

2.4. Dovetail Genomics Omni-C

The Dovetail Omni-C libraries were generated using Dovetail Genomics Omni-C kit (Dovetail Genomics, DG-REF-001, DG-REF-002, DG-LIB-001) according to the manufacturer's instructions. Single indexed primers supplied in the Dovetail Primer Set Module (Dovetail Genomics, DG-PRS-001) were used for the indexing step. The QC sequencing analysis was performed for each Omni-C library as described in the QC analysis pipeline (https://github.com/dovetail-genomics/omni-c_qc) before proceeding to the target capture (Section 2.6).

2.5. Target Enrichment Array Design for rCHi-C

This project required me to design a comprehensive array, that would include as many potentially causal variants as possible (limited by the maximum array size of 24 Mb) and would be compatible with different CHi-C protocols. Final array regions were still defined on the basis of HindIII restriction sites, as this is the least frequent cutter enzyme between the three protocols. As a result, capturing the entire fragments of interest (instead of capturing just the ends of the fragments) allows for the Arima RE mix and Dovetail DNase I restriction sites, providing extensive coverage of each locus.

Details of the array design can be found in Figure 2.4. Briefly, 211 index SNPs associated with breast cancer (N = 190), mammographic density (N = 11) or breast size (N = 10) were selected based on three studies[67, 74, 107]. All variants correlated with the index SNPs ($r^2 \geq 0.6$; N = 12,284) were identified based on 1000 Genomes Phase 3 data for 181 individuals of two European populations (CEU, GBR). The R package 'proxysnps' (https://github.com/slowkow/proxysnps) was used to remove non-founder samples from the analysis and select these correlated SNPs. The number of SNPs was reduced to 12,097 SNPs after excluding those with low minor allele frequency (MAF < 0.01). Seven index SNPs were also found to have MAF < 0.01; these SNPs were retained and treated separately for the purpose of array design (described in Figure 2.4). For the preliminary design, capture regions were defined as the regions that included all correlated SNPs ($r^2 \geq 0.6$) for each of the index SNPs. This resulted in 25 'large' regions (> 250 kb in size) that were treated separately for array design purpose. For these, I used a higher threshold ($r^2 \geq 0.9$) of correlation to reduce their size (allowing me to retain 18 regions on the array); the other seven regions remained intractable and were excluded.

The other 179 regions ($r^2 \geq 0.6$; < 250 kb) were cross-checked with 142 fine-scale mapping regions (published by BCAC), covering 4,453 credible variants[66]. Out of 179 regions, 66 did not overlap with any credible variants, 94 covered all credible variants at the given region and 19 partially overlapped with the credible variants.

For the final array I made the following pragmatic decision – for the regions where less than 30 credible variants were missing, each missing variant was captured separately, while for the regions where there were more than 30 variants missing, the missing variants were not included. Each region was then mapped to corresponding HindIII fragments

with 500 bp added at either end, to take account of the fragment size generated by shearing, and 120-mer RNA baits were designed using Agilent eArray software (https://earray.chem.agilent.com/suredesign/). A list of regions targeted by the array can be found in Appendix A.

The focus of my thesis is the breast cancer risk regions; MD and breast size loci will be the subject of a subsequent analysis. The final array targeted 183 'strong-evidence' breast cancer risk signals across 122 regions. Excluded signals can be found in Table 2.7.

Fine-mapping region	Signal	Index SNP	CCVs
chr5:50133661-50977284*	Signal 1	rs373575834	4
	Signal 2	rs3846498	93
chr5:81632442-82742227*	Signal 1	rs2059891	23
chr6:28454660-29458443*	Signal 1	rs79309050	36
chr7:91501305-92552283*	Signal 1	rs7785971	318
chr10:120834389-122089809	Signal 1	rs35054928	1
	Signal 5	rs7899765	1
chr12:27486913-28881482	Signal 2	rs1600346	375
chr17:45675102-46675102*	Signal 1	rs62070949	2277
chr18:26252512-27495432	Signal 4	rs180952292	1
chr19:18937437-19937437*	Signal 1	rs1469713	162
chr22:45387417-46387400*	Signal 1	rs11704298	34
	signal 2	rs184070480	1

Table 2.7: Breast cancer risk signals that were not targeted by the rCHi-C array. 13 'strong-evidence' signals (from Fachal et al. 2020) that were not targeted by my rCHi-C array. In cases where a fine-mapping region is marked with (*), the entire region was excluded from the array design. CCVs – number of CCVs reported at the signal. Fine-mapping regions are in GRCh38/hg38.

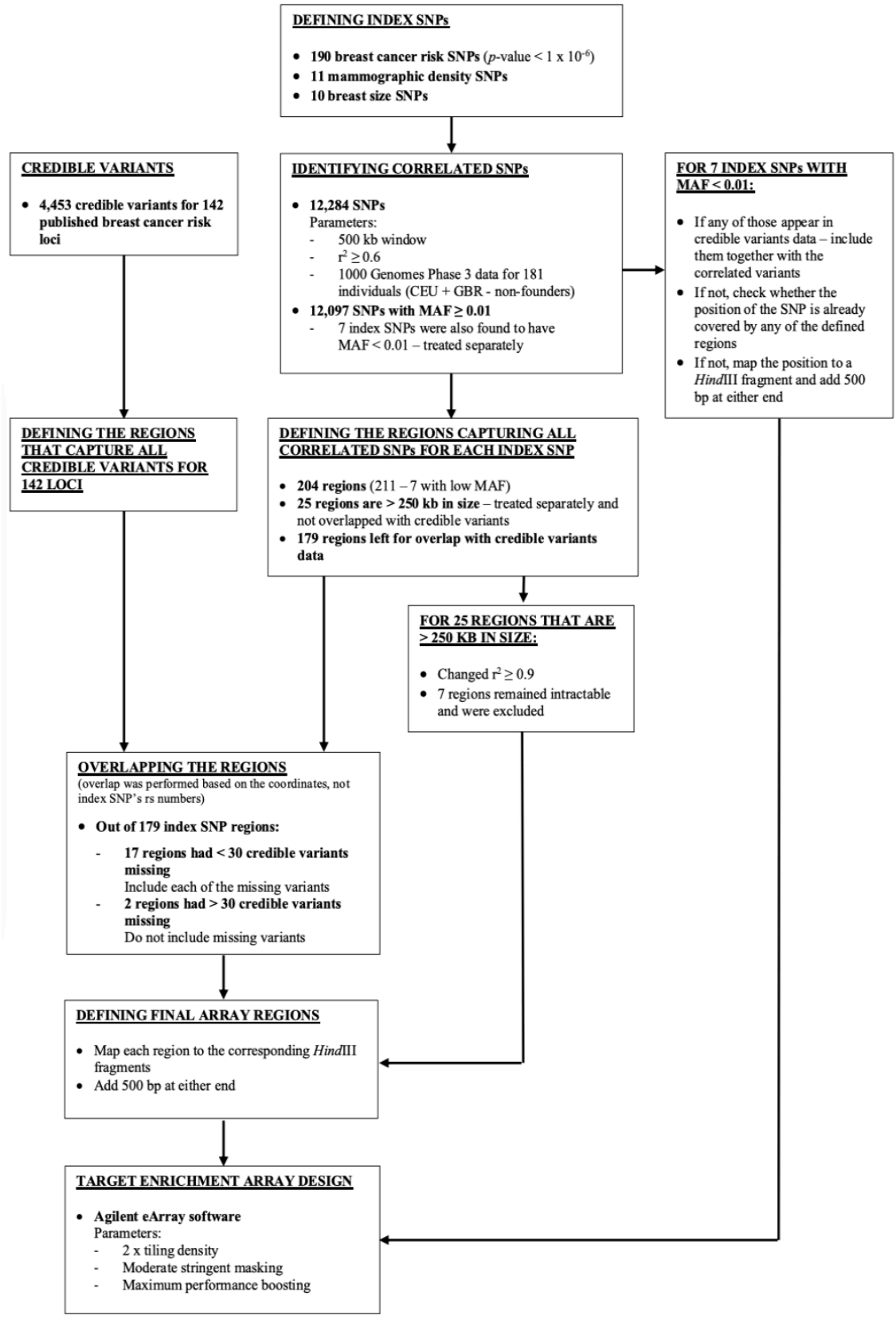

Figure 2.4: Target enrichment array design.

2.6. Target Sequence Capture

2.6.1. Region CHi-C

Hi-C libraries generated by all three protocols (Standard Hi-C – Section 2.2, Arima Genomics Hi-C – Section 2.3, Dovetail Genomics Omni-C – Section 2.4) were subjected to target enrichment using Agilent SureSelect kit (Agilent, 5190-9685, 5190-9684). To prepare libraries for hybridisation, 500 – 750 ng of each library were transferred to a new tube, dehydrated in a vacuum concentrator using low heat (30°C) and resuspended in 3.4 µl nuclease-free water.

Per reaction, 13 µl hybridisation buffer (6.63 µl SureSelect Hyb 1, 0.27 µl SureSelect Hyb 2, 2.65 µl SureSelect Hyb 3 and 3.45 µl SureSelect Hyb 4) and 5.6 µl block mix were prepared. Components of the block mix vary depending on the protocol used for the Hi-C library generation. For the Standard Hi-C protocol, the block mix consisted of 2.5 µl SureSelect Indexing Block 1, 2.5 µl SureSelect Block 2 and 0.6 µl SureSelect ILM Indexing Block 3. For the Arima Hi-C and Dovetail Omni-C libraries, 2 µl xGen Universal Blockers TS, 16 rxn (IDT, 1075474) and 5 µl SureSelect XT HS and XT Low Input Blocker Mix were used, respectively, topped up to 5.6 µl with nuclease-free water. Hybridisation buffer was incubated at 65°C for 5 minutes followed by equilibration at RT, while block mix was kept on ice at all times before being added to the library.

After addition of the block mix, libraries were incubated at 95°C for 5 minutes and then held at 65°C for at least 5 more minutes. In this time, capture library hybridisation mix was prepared by combining 13 µl hybridisation buffer, 2 µl 25% RNase block and 5 µl SureSelect Capture Library.

Keeping reaction tubes at 65°C, 20 µl capture library hybridisation mix was added to each library and mixed by pipetting. Samples were re-sealed and incubated for 24 hours at 65°C with lid temperature of 105°C.

To capture hybridised DNA, Dynabeads MyOne Streptavidin T1 (Invitrogen, 65601) beads were used. Per reaction, 50 µl beads were prepared by washing three times with 200 µl SureSelect Binding Buffer, reclaiming the beads against the MPS and resuspending in 200 µl SureSelect Binding Buffer. Next, the entire volume of

hybridisation reaction was transferred directly from the 65°C block to the resuspended beads, followed by a 30-minute incubation at RT on a rotator.

DNA-bound beads were reclaimed against the MPS, resuspended in 200 µl SureSelect Wash Buffer 1 and incubated at RT for 15 minutes. After being reclaimed against the MPS, beads were washed three times with 200 µl SureSelect Wash Buffer 2. For each wash, beads were resuspended in pre-warmed to 65°C buffer, incubated at 65°C for 10 minutes and reclaimed against the MPS. After removal of the final wash, beads were resuspended in 50 µl nuclease-free water.

In preparation for sequencing, captured libraries were subjected to post-hybridisation PCR. Since adapters and indexes used in the Standard Hi-C, Arima Genomics Hi-C and Dovetail Genomics Omni-C protocols vary, different PCR primers were used (Table 2.8). Libraries generated using the Standard Hi-C protocol are not indexed, so the unique termini (P5 and P7 sequences) that are required for binding to the flow cell were added at this stage. Therefore, Posthyb Fw Standard and Posthyb Rv Standard primers contain sequences that are partially complementary to the adapters used and the P5/P7 sequences. Libraries generated by the Arima Genomics protocol are dual-indexed. i5 and i7 indexes and P5/P7 sequences were introduced in the adapter annealing step, so pre-hybridisation primers were also used in the post-hybridisation PCR amplification step. Libraries generated by the Dovetail Genomics Omni-C protocol are single-indexed. Variable i7 index and universal i5 sequence were introduced during the indexing step (equivalent to the pre-hybridisation PCR), so the same reverse primer was used for the Dovetail Genomics libraries as for the Arima libraries. Forward primer was re-designed to only contain sequence complementary to the P5 sequence.

Primer Name	Protocol	Sequence
Posthyb Fw Standard	Standard Hi-C	AATGATACGGCGACCACCGAGATCTACACTCT TTCCCTACACGACGCTCTTCCGATC*T
Posthyb Rv Standard	Standard Hi-C	CAAGCAGAAGACGGCATACGAGATCGGTCTC GGCATTCCTGCTGAACCGCTCTTCCGATC*T
Arima Fw	Arima Hi-C	AATGATACGGCGACCACCGAGATCTACA*C
Arima Rv	Arima Hi-C and Dovetail Omni-C	CAAGCAGAAGACGGCATACGAGA*T
Dovetail Fw	Dovetail Omni-C	AATGATACGGCGACCACCG*A

Table 2.8: Post-hybridisation PCR primers. * indicates phosphorothioate bond, resistant to exonuclease degradation. Red – P5 sequence; blue – P7 sequence.

Step	Temperature	Time	Cycles
1	98°C	30 sec	1
2	98°C	10 sec	4
3	65°C	30 sec	
4	72°C	30 sec	
5	72°C	5 min	1
6	4°C	∞	1

Table 2.9: Post-hybridisation PCR thermocycle conditions.

Multiple PCR reactions were set up for each library containing 2.5 µl on-bead DNA, 0.4 µl 25 µM forward and 0.4 µl 25 µM reverse primers, 12.5 µl NEBNext High-Fidelity 2X PCR Mastermix and 9.2 µl water). PCR Thermocycle conditions are shown in Table 2.9.

PCR product from these reactions was cleaned up using AMPure XP beads as described in Section 2.2.12. Library DNA quality was assessed using the Agilent 2100 Bioanalyzer instrument (High Sensitivity DNA Assay) before proceeding to next-generation sequencing (NGS).

2.6.2. Promoter CHi-C

Promoter CHi-C libraries were generated as a part of Beta-testing project of the Dovetail Genomics Human Pan Promoter Enrichment Kit. Since this kit incorporates target enrichment technology from Twist Bioscience, target sequence capture was performed as described in the Twist Target Enrichment Protocol. The Dovetail Genomics Custom Panel supplied with this kit covered 84,643 promoters associated with 27,375 coding and non-coding genes. Since the total panel size is between 10 and 50 Mb, eight PCR cycles were used for the post-capture PCR amplification. Additionally, KAPA HiFi HotStart ReadyMix was replaced by Capture Amplification Mix supplied by Dovetail Genomics.

2.7. Sequencing

Region CHi-C libraries were sequenced at the ICR Genomics Facility on an Illumina NovaSeq 6000 System generating 100 bp paired-end reads. Promoter CHi-C libraries were sequenced on an Illumina HiSeq X Ten System (2 x 150 bp) by Dovetail Genomics. Libraries generated using the Standard Hi-C protocol were sequenced to generate ~ 40 Gb of raw data per library (in a single run). The GS2 Arima library was sequenced to

achieve ~ 100 Gb of data (in a single run). All other libraries were sequenced to a total of ~ 100 Gb of raw data per library in two separate 50 Gb runs. Sequencing data were output in fastq format.

2.8. Data Processing

Harriet Kemp, Andrea Gillespie and Syed Haider carried out the bioinformatics required to convert the raw sequencing data from the CHi-C experiments into a set of statistically significant interaction peaks.

For libraries generated using the Standard protocol, alignment of the fastq sequencing reads to the human reference genome (GRCh38/hg38) was performed using Bowtie2 aligner as a part of the Hi-C User Pipeline (HiCUP). Full details of this pipeline are available at http://www.bioinformatics.babraham.ac.uk/. HiCUP was also used to remove experimental artefacts and PCR duplicates.

Libraries generated using the Arima Genomics Hi-C and Dovetail Genomics Omni-C kits were processed using the Arima (https://github.com/ArimaGenomics/mapping_pipeline) and Dovetail (https://omni-c.readthedocs.io/en/latest/) pipelines, respectively. Briefly, in both pipelines, fastq sequencing reads were aligned to the human reference genome using the BWA aligner followed by the alignment quality and PCR duplicates filtering.

Removal of off-target di-tags (defined as di-tags where neither end mapped to one of the capture regions) and subsequent interaction peak calling was performed using CHiCANE[108]. Data from single library replicates were used for cell line rCHi-C analysis, while data from two technical replicates were pooled together for the pCHi-C analysis. Because pCHi-C array is approximately twice the size of my rCHi-C array (19,144 kb vs. 9,782 kb), this resulted in a similar numbers of on-target pairs per kb of array. For rCHi-C analysis in primary cells, data from two biological replicates were combined.

Interaction peaks in Standard libraries were called using individual HindIII fragments as the unit of analysis, while Arima and Dovetail libraries were called using both 2kb- and 5kb-binned data. The Dovetail Genomics protocol was used to generate both rCHi-C and pCHi-C. For the downstream analysis of pCHi-C data, Dovetail processing pipeline allocates gene promoters into 39,825 smart bins. Most smart bins capture promoter(s) of

a single gene, but in some smart bins (N=2,905) the proximity of gene promoters is such that a single smart bin captures multiple promoters. In addition, a subset of genes (N=9,104) occupies more than one smart bin. Therefore, to facilitate integration analysis (rCHi-C + pCHi-C), data obtained from rCHi-C libraries generated using the Dovetail Genomics protocol were binned in a promoter aware manner (using smart bins). Only significant CHiCANE interaction peaks (q-value ≤ 0.1) were considered for downstream analysis.

Aligning of interaction peaks with annotated gene promoters and CCVs, and the analysis to determine whether third-party bins are enriched for CTCF and H3K27ac binding was carried out by Harriet Kemp. Odds ratios and p values were calculated using a Fisher's exact test.

The CTCF datasets (all cell types, all breast, primary breast) that are shown in a subset of figures in which looping interactions are aligned with other relevant features and that were used for the third-party bins' enrichment analysis were compiled by Andrea Gillespie. These datasets are combinations of ChIP-seq data available from the Encyclopedia of DNA Elements (ENCODE)[109]. Data were collated using all available CTCF ChIP-seq datasets as of 31/01/2021. All cell types – consensus peaks identified in at least 28 out of the 31 available cell types. All breast – consensus peaks identified in at least 8 out of 9 breast relevant cell types for which data were available (3 – breast epithelium, 1 – mammary epithelial cells, 1 – mammary fibroblasts, 4 – MCF-7). Primary breast – consensus peaks identified in both of the primary breast cell types that were available (1 – mammary epithelial cells, 1 – mammary fibroblasts).

3. Region Capture Hi-C in T-47D and GS2 cells

3.1. Summary sequencing statistics

Using three different protocols (Standard, Arima and Dovetail), rCHi-C libraries were generated in T-47D and GS2 cell lines. T-47D is an ER+, breast cancer cell line, that is well-studied and frequently used in breast cancer research; GS2 is a 'normal' immortalised breast fibroblast cell line, which has not been previously used for chromosome conformation capture. Table 3.1 shows sequencing statistics for these libraries. Two key metrics that allow assessment of library quality and comparison of the methods are: the proportion of unique read pairs and the proportion of on-target di-tags (defined as read pairs for which at least one end colocalised with a capture panel probe).

Comparing these metrics between the protocols, the proportion of unique pairs varied from 53% to 77%, with the highest proportions observed in the Arima libraries. The proportion of on-target pairs varied from 10% to 38%, with the highest proportion observed in the T-47D Standard library and the lowest proportion in the T-47D Arima library. Interestingly, the proportion of on-target pairs appeared to be higher in libraries with lower absolute numbers of unique pairs.

As mentioned previously (Section 1.4.1.1), very short-range interactions often represent dangling ends flanking an undigested restriction fragment (partial digest). Such invalid interactions increase in frequency with decreasing restriction fragment size and are challenging to identify computationally. Accordingly, another useful metric to look at is the proportion of read pairs where the distance between interacting fragments (when mapped back to the reference genome) is less than 1 kb, since these are unlikely to represent 'true' interactions and can be considered wasted sequencing. Since in the Standard libraries the average size of restriction fragments is larger (3 – 4 kb), partial digest dangling ends are less likely to occur in these libraries and short-range interactions would be defined more accurately as those where the distance between interacting fragments is less than 10 kb. Therefore, as expected, the Standard libraries had the lowest proportion of short-range interactions – 6.5% and 5.6% (0.5% and 0.6% *cis* ≤ 1kb; 6.0% and 5.1% *cis* 1 kb – 10kb), while the Dovetail libraries had the highest (23.5% and 15.8%).

Method	Cell line	Total pairs	Unique pairs		On-target pairs		*cis* pairs		*cis* ≤ 1kb		*cis* 1kb - 10kb		*cis* 10kb - 1Mb		*cis* > 1Mb	
Dovetail	T-47D	368,720,573	214,685,595	58%	36,985,965	17%	31,196,050	84%	7,340,953	24%	4,555,742	15%	12,536,036	40%	6,763,319	22%
Dovetail	GS2	441,740,325	250,167,558	57%	43,554,308	17%	34,370,927	79%	5,432,680	16%	5,163,740	15%	15,121,596	44%	8,652,911	25%
Standard	T-47D	209,109,121	110,241,257	53%	41,631,203	38%	33,209,433	80%	168,037	0.5%	2,002,164	6%	18,739,495	56%	12,299,737	37%
Standard	GS2	312,566,378	213,308,313	68%	39,626,364	19%	28,653,821	72%	161,536	0.6%	1,453,290	5%	14,737,517	51%	12,301,478	43%
Arima	T-47D	421,801,216	257,001,042	61%	25,115,067	10%	20,599,509	82%	1,520,671	7%	3,751,012	18%	10,995,365	53%	4,332,461	21%
Arima	GS2	239,860,624	184,105,614	77%	39,266,593	21%	33,078,541	84%	3,433,968	10%	7,867,403	24%	16,188,027	49%	5,589,143	17%

Table 3.1: Summary sequencing statistics for T-47D and GS2 rCHi-C libraries. Summary sequencing statistics for the rCHi-C libraries generated in T-47D (ER+, breast cancer cell line) and GS2 ('normal' immortalised breast fibroblast cell line) cell lines using the Standard in-house Hi-C (Standard), the Arima Hi-C (Arima) and the Dovetail Genomics Omni-C (Dovetail) protocols. Total pairs – total number of read pairs where both ends aligned uniquely to the reference genome. On-target pairs – read pairs for which at least one end overlaps with a capture array probe (minimum overlap = 1 bp).

Method	Cell line	Bin size	Total IPs	*trans* IPs		*cis* < 1kb		*cis* 1kb - 10kb		*cis* 10kb - 100kb		*cis* 100kb - 1Mb		*cis* ≥ 1Mb	
Dovetail	T-47D	5kb	9,985	557	5.6%	0	0.00%	13	0.1%	224	2.2%	4,067	41%	5,124	51%
Dovetail	T-47D	2kb	12,885	24	0.2%	6	0.05%	69	0.5%	870	6.8%	11,150	87%	766	6%
Arima	T-47D	5kb	9,907	1,128	11.4%	0	0.00%	4	0.0%	140	1.4%	4,312	44%	4,323	44%
Arima	T-47D	2kb	15,002	168	1.1%	0	0.00%	21	0.1%	848	5.7%	13,398	89%	567	4%
Standard	T-47D	NA	18,407	3,716	20.2%	2	0.01%	19	0.1%	154	0.8%	5,839	32%	8,677	47%
Dovetail	GS2	5kb	12,059	68	0.6%	0	0.00%	6	0.0%	152	1.3%	4,945	41%	6,888	57%
Dovetail	GS2	2kb	18,342	47	0.3%	6	0.03%	18	0.1%	1,095	6.0%	15,750	86%	1,426	8%
Arima	GS2	5kb	12,645	142	1.1%	0	0.00%	7	0.1%	127	1.0%	5,675	45%	6,694	53%
Arima	GS2	2kb	24,845	107	0.4%	0	0.00%	33	0.1%	1,165	4.7%	21,343	86%	2,197	9%
Standard	GS2	NA	11,121	79	0.7%	0	0.00%	10	0.1%	49	0.4%	5,063	46%	5,920	53%

Table 3.2: Interaction peak calling statistics for T-47D and GS2 rCHi-C libraries. A breakdown of significant (q-value ≤ 0.1) interaction peaks (IPs) called in the rCHi-C libraries generated in two cell lines using three different protocols and called using CHiCANE is shown. Standard libraries were called using individual HindIII fragments as the unit of analysis, while Arima and Dovetail libraries were called using both 2kb- and 5kb-binned data.

3.2. Interaction peak calling

Interaction peaks were called using CHiCANE[108]– an in-house pipeline developed specifically for the analysis of rCHi-C data. Standard libraries were called using individual HindIII fragments as the unit of analysis, while Arima and Dovetail libraries were called using both 2kb- and 5kb-binned data. For simplicity, I will refer to both HindIII fragments and Arima/Dovetail bins as 'bins' for the rest of the thesis. Only significant CHiCANE interaction peaks (q-value ≤ 0.1) were considered for further analysis.

The number of interaction peaks called per library varied from 9,907 to 24,845. Table 3.2 shows the breakdown of these interaction peaks. Interestingly, the proportion of *trans* interaction peaks was higher in the T-47D libraries (0.2% to 20.2%) than in GS2 libraries (0.3% to 1.1%), potentially because T-47D cells are highly re-arranged cancer cells. Higher *trans* proportions were consistently observed in the Standard and 5kb-binned Arima and Dovetail libraries, than in the 2kb-binned libraries. Additionally, the Standard and 5kb-binned libraries had much higher proportions of *cis* interaction peaks in the ≥ 1 Mb range, which is unlikely to be the most relevant range for functional follow up of GWAS risk loci[110, 111].

3.3. Overview of all interaction peaks

In a recent fine-scale mapping analysis, the BCAC identified 7,394 CCVs within 196 'strong-evidence' ($p < 1 \times 10^{-6}$) signals across 129 genomic regions[67]. It is generally accepted that a large proportion of functional non-coding GWAS variants influence breast cancer risk by disrupting regulatory elements that mediate expression of target gene(s)[78, 80]. Therefore, the rationale for using 3C-based technologies for annotation of risk loci is that identifying regulatory interactions between a genomic region that harbours a CCV and a genomic region that colocalises with a gene promoter could prioritise a subset of CCVs and putative target genes for in-depth functional studies.

Accordingly, rCHi-C data was mapped to: (i) 84,643 promoters associated with 27,375 coding and non-coding genes; (ii) 5,117 out of 7,394 CCVs reported by Fachal and colleagues[67] (the remaining 2,277 CCVs were accounted for by one signal, resulting from

strong LD with a CNV, and, therefore, were excluded from further analysis). The results are illustrated in Table 3.3.

The proportion of unique gene-containing bins (defined as bins that colocalised with gene promoter(s) and formed at least one interaction peak) out of the total unique bins in each given dataset varied from 2.3% to 9.7%. There were 113 to 453 unique genes that participated in 345 to 4,870 interaction peaks with a median of 2 to 4 interaction peaks per individual gene. The proportion of unique CCV-containing bins (defined as the bins that colocalised with at least one CCV and formed at least one interaction peak) out of the total unique bins in each given dataset varied from 5.1% to 7.7%. These CCV-containing bins harboured a total of 684 to 1,190 unique CCVs (13.4% to 23.3% out of 5,117 CCVs) that participated in 3,097 to 7,463 interaction peaks. Interestingly, the 5kb-binned Dovetail libraries had the lowest absolute numbers (and proportions) of gene-containing bins. In addition, when comparing 5kb and 2kb Dovetail data, the difference between numbers of gene-containing bins was much more pronounced than between numbers of CCV-containing bins.

3.4. Direct interaction peaks

The overall aim of my CHi-C experiments is to prioritise a set of CCVs and putative target genes that warrant in-depth functional follow up. In order to do this, I focused on the subset of interaction peaks in which a bin colocalising with a gene promoter formed a direct interaction with a CCV-containing bin. Table 3.4 shows a summary of these direct interaction peaks. The number of direct interaction peaks varied from 47 to 678. They involved 27 to 272 unique gene-containing bins harbouring a total of 32 to 217 ‘direct’ genes (1 to 3 genes per bin). Overall, 105 to 616 CCVs were involved in direct interaction peaks (‘direct’ CCVs). The largest range of CCVs per bin (1 to 40) was observed in Standard libraries, reflecting the larger average size of a restriction fragment.

Comparing ‘direct’ genes across differentially binned Arima and Dovetail datasets (Figure 3.1), almost three times as many individual genes were involved in direct interaction peaks in the 2kb Dovetail data versus 5kb data. For Arima, numbers were more similar between the datasets, with more individual genes found in the 5kb Arima data versus 2kb Arima data (except that more gene-containing bins were identified in the 2kb Arima GS2 data).

Method	Cell line	Bin size	Total IPs	Total unique bins in a dataset	Total unique gene-containing bins		Total unique genes	Number of IPs involving gene-containing bins	IPs per unique gene (median + range)	Total unique CCV-containing bins		Total unique CCVs (out of 5,117)		Number of IPs involving CCV-containing bins
Dovetail	T-47D	5kb	9,985	4,744	108	2.3%	113	345	2 (1 to 29)	322	6.8%	886	17.3%	3,658
Dovetail	T-47D	2kb	12,885	5,844	313	5.4%	307	1,529	2 (1 to 94)	417	7.1%	749	14.6%	3,097
Arima	T-47D	5kb	9,907	4,108	378	9.2%	311	1,978	3 (1 to 229)	295	7.2%	824	16.1%	3,571
Arima	T-47D	2kb	15,002	5,506	423	7.7%	279	2,122	3 (1 to 84)	363	6.6%	684	13.4%	3,599
Standard	T-47D	NA	18,407	7,113	585	8.2%	453	3,136	2 (1 to 224)	365	5.1%	1,085	21.2%	6,901
Dovetail	GS2	5kb	12,059	4,725	120	2.5%	120	414	2 (1 to 24)	342	7.2%	956	18.7%	5,611
Dovetail	GS2	2kb	18,342	7,075	379	5.4%	359	2,565	4 (1 to 94)	504	7.1%	946	18.5%	5,633
Arima	GS2	5kb	12,645	5,063	493	9.7%	383	2,403	3 (1 to 108)	388	7.7%	1,190	23.3%	5,763
Arima	GS2	2kb	24,845	9,018	691	7.7%	410	4,870	4 (1 to 171)	548	6.1%	1,084	21.2%	7,463
Standard	GS2	NA	11,121	5,688	471	8.3%	363	2,082	2 (1 to 346)	292	5.1%	956	18.7%	5,028

Table 3.3: Summary of T-47D and GS2 interaction peaks for which the interacting fragments colocalised with: (i) an annotated RefSeq gene promoter; (ii) one or more CCVs selected by the BCAC fine-scale mapping analysis.

Method	Cell line	Bin size	Total direct IPs	Unique gene-containing bins	Unique genes	Unique CCV-containing bins	Unique CCVs	CCVs per bin
Dovetail	T-47D	5kb	47	27	32	37	105	1 to 8
Dovetail	T-47D	2kb	192	87	95	133	226	1 to 8
Arima	T-47D	5kb	331	144	124	139	395	1 to 19
Arima	T-47D	2kb	264	120	82	116	210	1 to 8
Standard	T-47D	NA	530	272	217	157	495	1 to 40
Dovetail	GS2	5kb	108	52	53	53	134	1 to 13
Dovetail	GS2	2kb	400	157	145	219	413	1 to 10
Arima	GS2	5kb	603	239	184	191	616	1 to 19
Arima	GS2	2kb	678	259	156	242	474	1 to 16
Standard	GS2	NA	423	227	177	141	504	1 to 40

Table 3.4: Summary of direct interaction peaks called in T-47D and GS2 rCHi-C libraries. Direct interaction peaks – interaction peaks in which a bin colocalising with a gene promoter forms a direct interaction with a CCV-containing bin.

In terms of similarity, 22 out of 32 genes found in the 5kb Dovetail T-47D dataset appeared in the 2kb dataset, and 42 out of 53 genes found in the 5kb GS2 dataset were in the 2kb dataset. For Arima, 55 out of 82 genes found in the 2kb T-47D dataset appeared in the 5kb dataset, and 104 out of 156 genes found in the 2kb GS2 dataset were in the 5kb dataset.

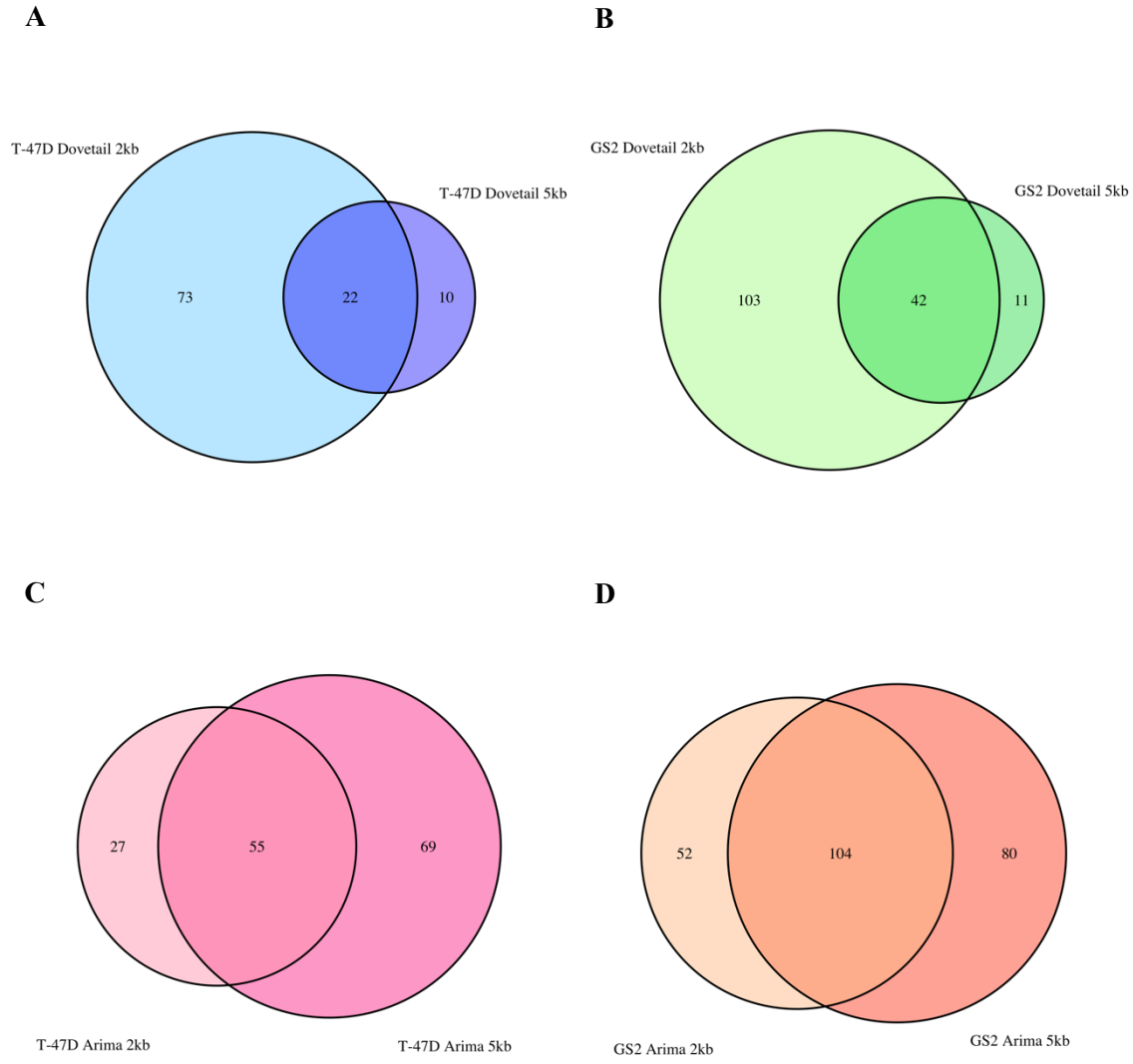

Figure 3.1: Venn diagrams illustrating the overlap between 'direct' genes identified in the 2kb- and 5kb-binned Dovetail and Arima rCHi-C libraries generated in T-47D and GS2 cells. (A) T-47D Dovetail 5kb and T-47D Dovetail 2kb rCHi-C datasets; (B) GS2 Dovetail 5kb and GS2 Dovetail 2kb rCHi-C datasets; (C) T-47D Arima 5kb and T-47D Arima 2kb rCHi-C datasets; (D) GS2 Arima 5kb and GS2 Arima 2kb rCHi-C datasets.

Comparing across the protocols, on average, 'direct' genes and CCVs found in Dovetail libraries had higher overlap with those in Arima libraries rather than with those in Standard libraries (Figure 3.2 and Figure 3.3). Additionally, there was higher overlap between 'direct' genes and CCVs between the Arima and Standard libraries, than between the Dovetail and Standard libraries.

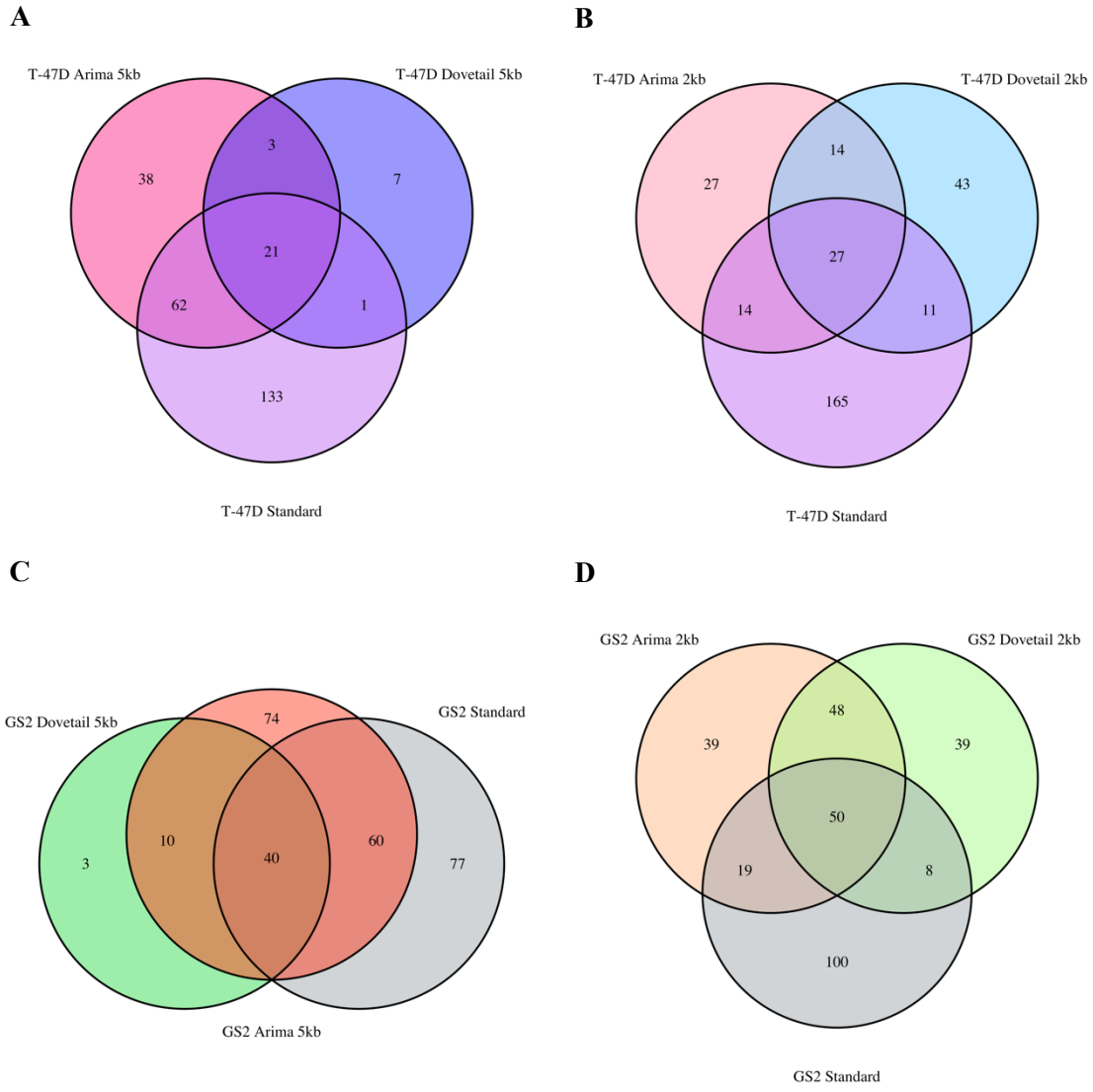

Figure 3.2: Venn diagrams illustrating the overlap between 'direct' genes identified in different T-47D and GS2 rCHi-C datasets. (A) T-47D Arima 5kb, T-47D Dovetail 5kb and T-47D Standard rCHi-C datasets; (B) T-47D Arima 2kb, T-47D Dovetail 2kb and T-47D Standard rCHi-C datasets; (C) GS2 Arima 5kb, GS2 Dovetail 5kb and GS2 Standard rCHi-C datasets; (D) GS2 Arima 2kb, GS2 Dovetail 2kb and GS2 Standard rCHi-C datasets.

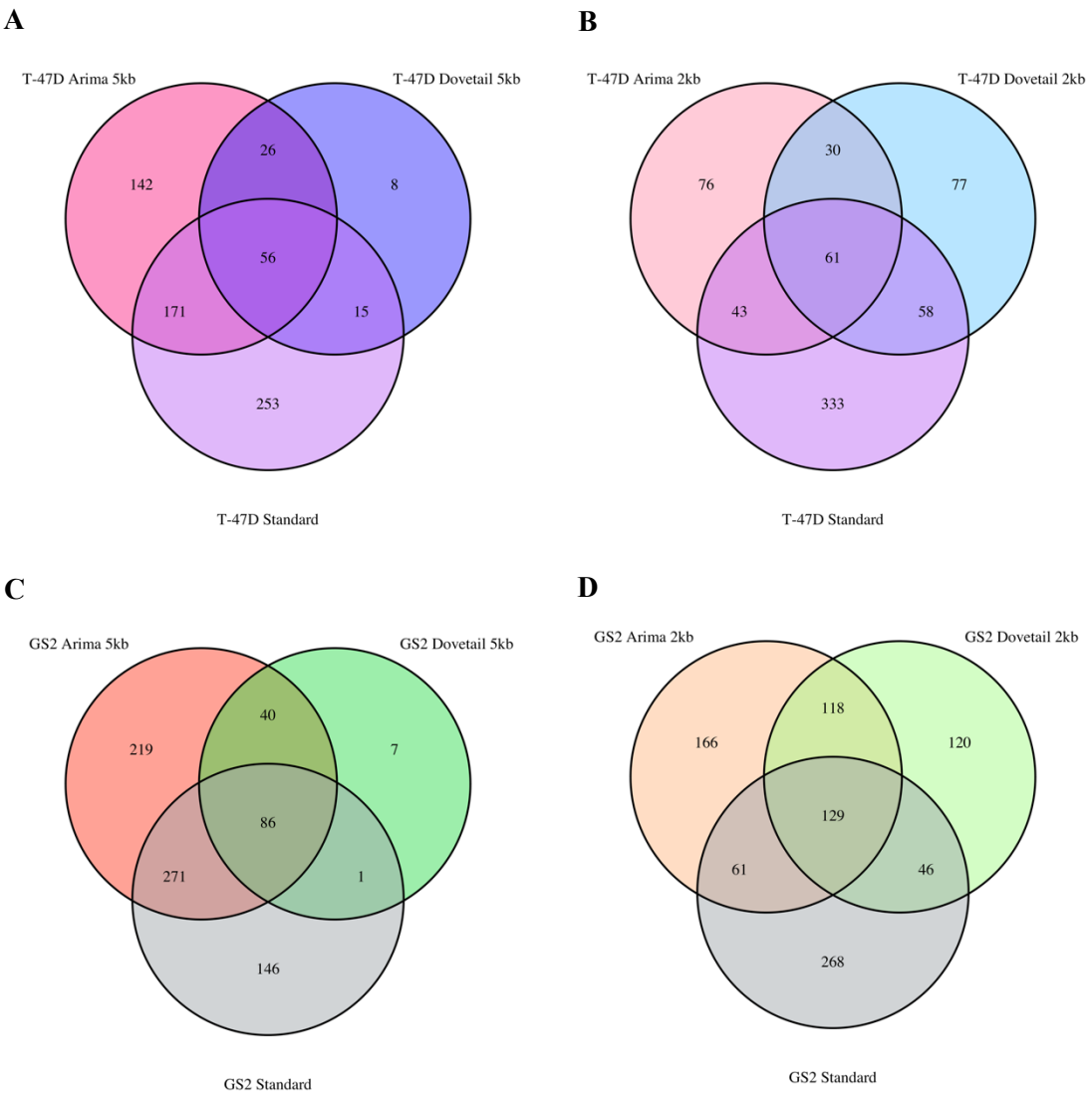

Figure 3.3: Venn diagrams illustrating the overlap between 'direct' CCVs identified in different T-47D and GS2 rCHi-C datasets. (A) T-47D Arima 5kb, T-47D Dovetail 5kb and T-47D Standard rCHi-C datasets; (B) T-47D Arima 2kb, T-47D Dovetail 2kb and T-47D Standard rCHi-C datasets; (C) GS2 Arima 5kb, GS2 Dovetail 5kb and GS2 Standard rCHi-C datasets; (D) GS2 Arima 2kb, GS2 Dovetail 2kb and GS2 Standard rCHi-C datasets.

Generally, the largest numbers of 'direct' genes and CCVs that were called in a single dataset (referred to hereafter as non-replicated genes and CCVs) were found in libraries generated using the Standard protocol. The only exception was in the 5kb-binned GS2 data, where there was a similar number of non-replicated genes and much larger number of non-replicated CCVs in the Arima library. As mentioned earlier, the Standard and 5kb-binned libraries had a higher proportion of *trans* and longer-range *cis* interaction peaks. To investigate whether this trend could be (at least partially) explained by these differences, I looked into the distribution of direct interaction peaks (Table 3.5).

Cell line	T-47D			GS2		
Dataset	**Total direct IPs**	**Very long-range (%)**	***trans (%)***	**Total direct IPs**	**Very long-range (%)**	***trans (%)***
2kb Dovetail	192	0	0	400	0	0
2kb Arima	264	0	0	678	2 (0.3%)	12 (1.8%)
5kb Dovetail	47	3 (6.4%)	0	108	2 (1.9%)	0
5kb Arima	331	31 (9.4%)	9 (2.7%)	603	63 (10.4%)	12 (2.0%)
Standard	530	104 (19.6%)	49 (9.2%)	423	44 (10.4%)	0

Table 3.5: Very long-range and *trans* direct interaction peaks in T-47D and GS2 rCHi-C libraries. Very long-range (here defined as *cis* > 2 Mb) and *trans* direct IPs called in T-47D and GS2 rCHi-C libraries.

In T-47D, the largest proportions of very long-range (19.6%) and *trans* (9.2%) interaction peaks were observed in the Standard library. Comparing to the 2kb-binned datasets, 82 out of 165 non-replicated genes (49.7%) and 131 out of 333 non-replicated CCVs (39.3%) found in the Standard library are explained by these 153 very long-range or *trans* interaction peaks. When comparing to the 5kb datasets, 70 out of 133 non-replicated genes (52.6%) and 67 out of 253 non-replicated CCVs (26.5%) came from these interaction peaks.

In GS2, 44 (10.4%) direct interaction peaks identified in the Standard library were very long-range. Comparing to the 2kb datasets, 31 out of 100 non-replicated genes (31%) and 79 out of 268 non-replicated CCVs (29.5%) in the Standard dataset are explained by these very long-range interaction peaks. When comparing to the 5kb datasets, 20 out of 77 non-replicated Standard genes (26%) and 20 out of 146 non-replicated Standard CCVs (13.7%) are explained by these interaction peaks. Among GS2 datasets, the highest proportion of very long-range and *trans* interaction peaks (12.4%) was observed in the 5kb Arima library, potentially explaining why it displayed a similar number of non-replicated genes and much larger number of non-replicated CCVs than the Standard library in Figures 3.3C and 3.4C. These potentially less informative interaction peaks included 14 out of 74 non-replicated genes (18.9%) and 29 out of 219 non-replicated CCVs (13.2%).

Overall, half (T-47D) and one-third (GS2) of 'direct' genes that were only identified in Standard protocol libraries mapped to very long-range or *trans* interaction peaks. These interaction peaks explain a lower proportion of non-replicated Standard CCVs, potentially because at least a subset of the remaining non-replicated CCVs can be

accounted for by some large HindIII fragments that contain many CCVs. Indeed, in the Standard libraries, there were 1 to 40 individual CCVs per HindIII fragment, while in the Arima and Dovetail libraries this ranges from 1 to 8 to 1 to 19 CCVs per bin (Table 3.4). Gene-containing bins, in turn, contained 1 to 3 individual 'direct' genes, regardless of the protocol.

3.5. The 2q35 locus

Since the true number of causal variants and target genes underlying each association signal is unknown, it is difficult to assess the quality of data generated by different protocols by examining the numbers alone. Therefore, to assess the data to the best of my ability, I examined a breast cancer risk locus at 2q35 that has been extensively characterised by this lab and others[101, 112-114]. Briefly, fine-scale mapping of the 2q35 locus has defined three independent 'strong-evidence' signals annotated by rs4442975 (signal 1; 1 CCV), rs138522813 (signal 2; 5 CCVs) and rs5838651 (signal 3; 42 CCVs) (Figure 1.1). Follow up studies identified *IGFBP5* as the target gene, and rs4442975 and a structural variant esv3594306 (~ 1.4 kb deletion) as the likely causal variants at signals 1 and 2 respectively, with causal variant(s) at signal 3 remaining unknown.

The number of interaction peaks involving the *IGFBP5* promoter varied from 1 to 50 between the datasets. On average, more activity was observed in the 2q35 region in GS2 libraries rather than in T-47D libraries. Although the functional variant at signal 1 formed at least one interaction peak in eight out of ten datasets (not in the 5kb-binned Dovetail and Arima GS2 libraries), only four datasets (Standard GS2, 5kb Arima T-47D, 2kb Dovetail GS2 and T-47D) picked up a direct interaction between this variant and the *IGFBP5* promoter (Figure 3.4 and Figure 3.5)

Only 2kb-binned datasets identified any interaction peaks at signal 2 at all; and three of these datasets (2kb Arima GS2, 2kb Dovetail GS2 and T-47D) picked up a direct interaction peak between a bin containing esv3594306 (together with rs572022984) and the *IGFBP5* promoter. Interestingly, the 2kb-binned Arima GS2 library also picked up a direct interaction peak between esv3594306-containing bin and *AC007563.3* – a lncRNA that has not been studied in the context of cancer.

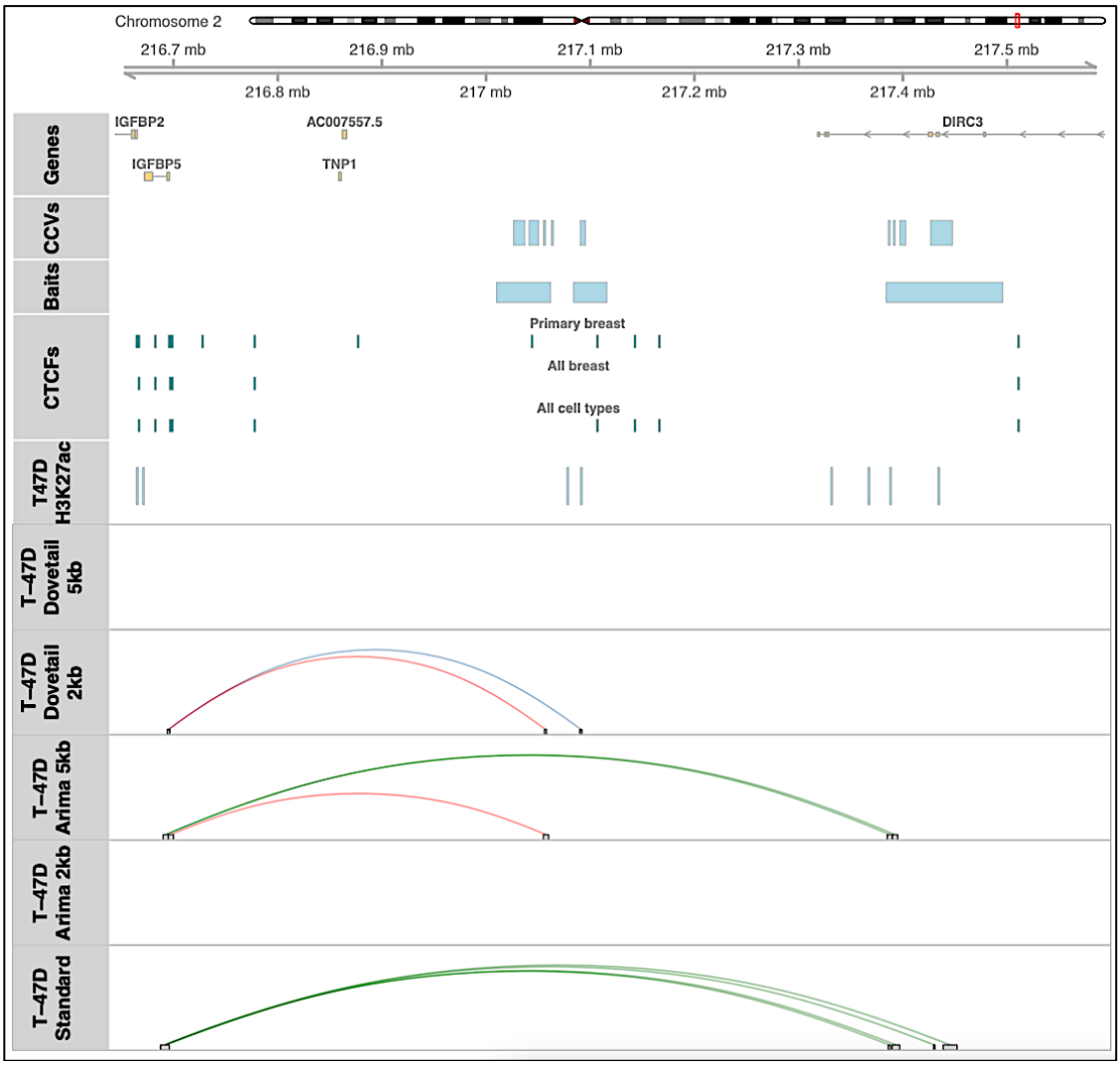

Figure 3.4: Direct interaction peaks at 2q35 in T-47D cells. Direct interaction peaks (shown in a looping format) at the 2q35 breast cancer risk locus (chr2:216,541,109-217,931,785 fine-mapping region, hg38) detected in T-47D rCHi-C data generated using the Standard, Arima and Dovetail protocols. Red loops – direct IPs that involved rs4442975 (signal 1). Blue loops – direct IPs that involved signal 2 CCVs. Green loops – direct IPs that involved signal 3 CCVs. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. Baits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). T47D H3K27ac – H3K27ac peaks identified from T-47D CUT&Tag data.

At signal 3, six datasets picked up at least one direct interaction peak involving the *IGFBP5* promoter (5kb Arima T-47D and GS2, Standard T-47D and GS2, GS2 Dovetail 5kb and 2kb), while the 2kb-binned Arima GS2 dataset picked up one direct interaction peak involving *AC007563.3*. Overall, the number of signal 3 CCVs that formed at least one direct interaction peak with the *IGFBP5* promoter varied from 2 to 42. rs6706673 formed at least one direct interaction peak with the *IGFBP5* promoter in four datasets (5kb Arima T-47D and GS2, Standard T-47D and GS2). rs6723847 formed at least one direct interaction peak in five datasets (5kb Arima T-47D and GS2, Standard T-47D and

GS2, 5kb Dovetail GS2). Often, however, due to the tight LD in that signal, there were multiple CCVs (2 to 27) per interacting bin. Although, the use of the 2kb-binned data as opposed to the 5kb-binned or Standard (HindIII) data broke down some of the large fragments, it remained difficult to confidently prioritise any of the variants.

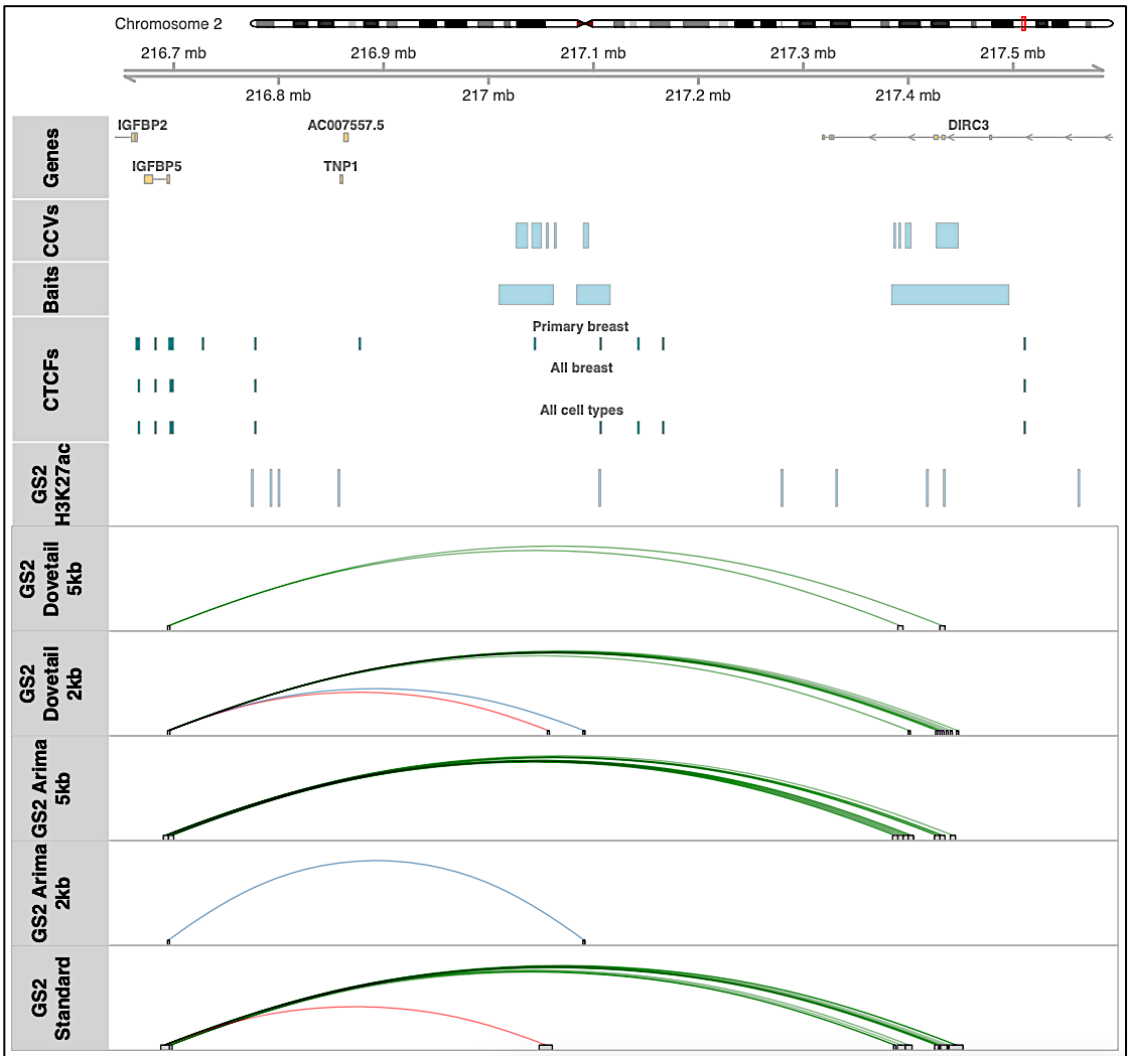

Figure 3.5: Direct interaction peaks at 2q35 in GS2 cells. Direct interaction peaks (shown in a looping format) at the 2q35 breast cancer risk locus (chr2:216,541,109-217,931,785 fine-mapping region, hg38) detected in GS2 rCHi-C data generated using the Standard, Arima and Dovetail protocols. Only direct interaction peaks involving the *IGFBP5* promoter are shown. GS2 Arima 2kb dataset also picked up one direct IP between esv3594306-containing bin (signal 2) and *AC007563.3*, and one between a bin containing rs3821098 and rs11693806 (signal 3) and *AC007563.3*. Red loops – direct IPs that involved rs4442975 (signal 1). Blue loops – direct IPs that involved signal 2 CCVs. Green loops – direct IPs that involved signal 3 CCVs. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. Baits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). GS2 H3K27ac – H3K27ac peaks identified from GS2 CUT&Tag data.

3.6. Prioritisation of putative target genes and CCVs

Overall, 2kb-binned as opposed to 5kb-binned Arima and Dovetail datasets, seem to perform better in many of the assessed metrics and provide more usable data. So, on the grounds that one of my goals was to increase library resolution, I focused on the 2kb-binned Arima and Dovetail datasets for the rest of the analysis.

'Direct' genes identified in the 2kb Dovetail, 2kb Arima and Standard datasets were mapped back to 129 'strong-evidence' breast cancer risk-associated regions (Table 3.6 and Table 3.7). Using T-47D datasets, at least one target gene was identified at 40 (Dovetail), 35 (Arima) and 42 (Standard) regions; while for GS2 these were found at 54 (Dovetail), 56 (Arima) and 42 (Standard) regions.

The true number of target genes at any locus is not known. It is likely that there will be more than one target gene at some loci, and it is possible that there are multiple genes contributing to the risk signal at several loci. Characterising a large number of genes and unravelling their combined effects would be challenging; pragmatically the number of genes that one can reasonably perform functional follow up on is limited to around 1 – 3 genes. Therefore, to evaluate the ability of different protocols to prioritise putative target genes that warrant in-depth functional follow up, I looked at numbers of putative target genes identified per region using each of the protocols (Table 3.8).

Fine-mapping region	T-47D Dovetail 2kb	T-47D Arima 2kb	T-47D Standard
chr1:9983762-11006158			RP4-635E18.9; TARDBP
chr1:45635245-46635245	NSUN4		
chr1:87191240-88191240		LINC01140	LMO4; LINC01140
chr1:113405767-114405767	RSBN1; HIPK1	HIPK1	
chr1:145625092-146010623*	NUDT17	NUDT17; PEX11B	
chr1:154676305-155678990	MTX1; THBS3; EFNA1; AC234582.1; AC234582.2; MUC1; RUSC1		SCNM1; TNFAIP8L2; EFNA3; SLC50A1; EFNA4; RP11-540D14.8; ADAM15; EFNA1
chr1:203332121-204332121	ETNK2; KISS1; SOX13; GOLT1A	ETNK2; GOLT1A; REN; SOX13; PLEKHA6; PIK3C2B; KISS1	ETNK2; GOLT1A; REN; SOX13; LINC00303; PLEKHA6; LINC00628
chr1:204049714-205049714	NFASC	NFASC; TMCC2	NFASC; PPP1R15B; TMCC2
chr1:216547232-217547232	ESRRG		
chr2:28447809-29447810			YPEL5; ALK; FAM179A
chr2:119987546-120988992			INHBB; AC073257.1; AC067960.1
chr2:171019711-172608243		DLX2; DLX2-AS1	DLX2; DLX2-AS1
chr2:216541109-217931785	IGFBP5		IGFBP5
chr3:4200592-5200591	BHLHE40; EDEM1	ARL8B; BHLHE40	SSUH2
chr3:26786474-28243756		AZI2; RP11-222K16.2; SLC4A7	NGLY1; OXSM; TOP2B; LINC00692; AZI2; AC114877.3; SLC4A7; RP11-222K16.1; CMC1
chr3:63456021-64482224	PRICKLE2; RP11-14D22.1; RP11-14D22.2	PRICKLE2; RP11-14D22.2; PSMD6	C3orf14; PRICKLE2; RP11-14D22.1; RP11-14D22.2
chr3:86488393-87488393		LINC00506	
chr3:140894017-141894017	ZBTB38		COL6A5; COL6A6; ATP2C1; ASTE1; NEK11
chr4:104647856-105935604	PPA2; GSTCD; INTS12	PPA2	PPA2
chr5:779675-1797374			MRFAP1L1; BLOC1S4; AC093323.3; RP11-539L10.3; PPP2R2C; RP11-539L10.2; S100P; KIAA0232; MAN2B2; MRFAP1; EGFLAM; C4orf50
chr5:15687249-16687419			RP1-251I12.1
chr5:44013202-45206396	PAIP1; RP11-53O19.3	RP11-53O19.2	NNT; TMEM267
chr5:56236057-57292056	CTC-236F12.4; IL6ST; ANKRD55; GPBP1	MAP3K1; CTC-236F12.4; C5orf67; RP11-155L15.1; IL6ST; ANKRD55; GPBP1	MAP3K1; AC008914.1; RP11-155L15.1; CDC20B; CCNO; IL6ST; CTC-236F12.4; ANKRD55; PLK2
chr5:58388234-59569743	PDE4D		
chr5:90993653-91993653	RP11-213H15.1; RP11-414H23.3; RP11-414H23.2; ADGRV1; LYSMD3	RP11-414H23.3; RP11-414H23.2	RP11-414H23.3; RP11-414H23.2; ADGRV1; NR2F1; SLF1; KIAA0825; FAM81B; CTC-458G6.2; CTC-529L17.2; POU5F2; RP11-185E12.2; MCTP1; RP11-133F8.2; CTC-529L17.1; RP11-348J24.1
chr5:132571366-133571367			C5orf15; WSPAR; TCF7
chr5:158303005-159317075	CTC-436K13.1; CTC-436K13.5		CTC-436K13.1; CTC-436K13.5
chr6:20121007-21121007	E2F3; SOX4	SOX4; RP11-204E9.1	SOX4; RP11-204E9.1; CASC15; NBAT1; LINC00581

chr6:80918669-82086234			RP11-244G12.1
chr6:129527974-130527974	RP1-69D17.4		
chr6:151097720-152615881		ESR1	
chr7:93984487-94984487	CALCR; PPP1R9A; PEG10; SGCE	PEG10; SGCE; ASB4	AC004012.1; ASB4; AC002429.5
chr8:36500965-37501668	ERLIN2		
chr8:74818066-76005702			PEX2; RP11-38H17.1; RP11-48D4.2; TPD52
chr8:100966731-101966731	GRHL2		
chr8:123097926-124097925		WDYHV1	
chr8:126412414-129029685	CASC11; MYC; PVT1	CASC11; MYC; LINC00824	LINC00977; RP11-26E5.1; CCDC26; RP11-419K12.1; RP11-136O12.2; LINC00976
chr10:8546150-9546150	GATA3; GATA3-AS1; RP11-379F12.4	GATA3; GATA3-AS1; RP11-379F12.4; TAF3	GATA3; GATA3-AS1; RP11-379F12.4; TAF3; RP5-1031D4.2; RP11-138I18.2; ITIH5; ITIH2
chr10:78581391-79627965	LINC00856		MRPS16; POLR3A; RPS24; RP11-157J13.1; LINC00856; RP11-90J7.3; MECOM; DLG5; DNAJC9; MYNN; SAMD7; AC074033.1; LRRC34; TERC; LRRIQ4; SEC62; GPR160; ACTRT3; RP11-90J7.2; PHC3; PRKCI
chr11:1377434-2421345	TNNT3		LSP1
chr11:65276356-66315595	BANF1; EIF1AD; KAT5; CFL1; MUS81		KAT5
chr11:107974789-108986410	ATM; NPAT; DDX10	DDX10	RP11-347E10.1
chr11:129082612-130091276	RP11-237N19.3	BARX2	RP11-673E11.2; FOXRED1; SRPRA; JAM3; FAM118B; RPUSD4; ESAM; VSIG2; RP11-469N6.3; ST3GAL4; PUS3; RP11-50B3.2; RP11-555G19.1; HEPN1; PKNOX2; TIRAP; CDON
chr12:13760997-14760997	ATF7IP	ATF7IP	
chr12:27486913-28881482		PTHLH	
chr12:114898717-115898717	TBX3; RP11-162N7.1	TBX3; RP11-162N7.1; RP4-601P9.2	TBX5; RP11-100F15.1; RP11-100F15.2; RBM19; TBX3; RP4-601P9.2; RP11-162N7.1; RP4-601P9.1; LHX5; RP11-25E2.1; WSB2; SDS; RP11-411G2.2; C12orf49; SDSL; TESC; TESC-AS1; RP11-497G19.2; RP11-989F5.3
chr12:119894342-120894343	MSI1; PLA2G1B; PXN	MSI1; DYNLL1; AL021546.6; COX6A1; PXN	PLA2G1B
chr14:36163563-37166547	PAX9; SLC25A21; MIPOL1	SLC25A21	SLC25A21
chr14:67650477-69067965	ZFP36L1		ZFP36L1
chr15:90465983-91465985		AC068831.16; VPS33B	FAM174B; LINC01578
chr16:3556787-4556787		ADCY9; LINC01569	
chr16:53267042-54321379	CRNDE; IRX5; IRX3; CTD-3032H12.2; RP11-434E6.4; FTO; CTD-3032H12.1	CRNDE; IRX5; CTD-3032H12.1; CTD-3032H12.2; IRX3; FTO; RPGRIP1L	CRNDE; IRX5; CTD-3032H12.1; CTD-3032H12.2; RP11-324D17.1; MMP2; IRX6; RP11-26L20.5; IRX3; RP11-212I21.2; AC007491.1; RP11-212I21.5; CTD-2302M15.1
chr16:80114430-81117200	CENPN; CMC2	CDYL2	WWOX; VAT1L; CLEC3A; NUDT7
chr16:86551631-87551631	FOXL1; FOXC2; MTHFSD	FOXL1; FOXC2; RP11-463O9.6	FOXL1; FOXC2; MTHFSD; RP11-118F19.1; GSE1

chr17:79794855-79816335*	CCDC40; TBC1D16	TBC1D16; RP11-353N14.4; CBX4	TBC1D16
chr19:12547463-14343759	IER2; STX10	DAND5	
chr19:29286822-30286822			PDCD5; ANKRD27; RGS9BP; CTD-2538C1.2; NUDT19; ZNF507; SLC7A9
chr21:14701662-15701664	AF127577.12; NRIP1	AJ009632.3	

Table 3.6: Distribution of T-47D 'direct' genes at 129 breast cancer risk regions. Out of 129 fine-mapping regions to which 196 'strong-evidence' breast cancer risk signals map, there were 40 at which 95 unique genes formed direct IPs in the T-47D Dovetail 2kb dataset, 35 at which 82 unique genes formed direct IPs in the T-47D Arima 2kb dataset and 42 at which 217 unique genes formed direct IPs in the T-47D Standard dataset. Regions that were not covered by the capture array as well as regions where no putative target genes were identified using any of the three datasets are not shown. Fine-mapping region coordinates are in GRCh38/hg38. (*) – there were several fine-mapping regions (originally defined in hg19) that when lifted over to hg38 were split or partially deleted in hg38. These regions were compiled manually to encompass all CCVs at each of the regions. Purple – genes that were called on the basis of *trans* interaction peaks.

Fine-mapping region	GS2 Dovetail 2kb	GS2 Arima 2kb	GS2 Standard
chr1:9983762-11006158	APITD1; APITD1-CORT		C1orf127; DRAXIN; MAD2L2; RP4-635E18.9; TARDBP; C1orf167
chr1:45635245-46635245	NSUN4		
chr1:87191240-88191240	LMO4; LINC01140	LMO4; LINC01140	LMO4; LINC01140
chr1:113405767-114405767	HIPK1	RSBN1; HIPK1	
chr1:145625092-146010623*	NUDT17; POLR3C; RNF115	NUDT17; POLR3C; RNF115	
chr1:154676305-155678990	MTX1; THBS3; RUSC1	MTX1; THBS3	CCT3; RHBG; LMNA
chr1:203332121-204332121	ETNK2; SOX13	SOX13; PIK3C2B; ETNK2; LINC00628	ETNK2; SOX13
chr1:204049714-205049714		NFASC	
chr2:18634525-19621042		LINC00954; OSR1	LINC00954
chr2:28447809-29447810	ALK	ALK; CLIP4	LINC01460; ALK; CLIP4; FAM179A; YPEL5
chr2:119987546-120988992		AC073257.2	AC073257.1; AC067960.1; NIFK; TSN
chr2:171019711-172608243	DCAF17; METTL8; DLX2; DLX2-AS1; SLC25A12; DYNC1I2; DLX1	CTB-25B13.5; DLX2; DLX2-AS1; HAT1; DYNC1I2; DLX1	CYBRD1; DLX2; DLX2-AS1; DYNC1I2; HAT1
chr2:172846180-173848166	LINC01305		SP9
chr2:200816524-201816524	CFLAR; FAM126B; NDUFB3	CASP10; CFLAR	
chr2:216541109-217931785	IGFBP5	IGFBP5; AC007563.3	IGFBP5
chr3:4200592-5200591	ARL8B	EDEM1	BHLHE40
chr3:26786474-28243756	NEK10	AZI2; CMC1; RP11-222K16.2; NEK10	AZI2; CMC1
chr3:63456021-64482224	ATXN7; THOC7; PRICKLE2; RP11-14D22.1; RP11-14D22.2	ATXN7; C3orf49; PRICKLE2; RP11-14D22.2; RP11-14D22.1; LINC00994	PRICKLE2; RP11-14D22.1; RP11-14D22.2
chr3:86488393-87488393	LINC00506	LINC00506	
chr3:140894017-141894017	ZBTB38		
chr4:82948971-83948971	NKX6-1	COQ2; NKX6-1	
chr4:104647856-105935604	TET2; PPA2; GSTCD; INTS12	PPA2	PPA2
chr5:44013202-45206396	NNT; PAIP1; FGF10	NNT	PAIP1
chr5:56236057-57292056	MAP3K1; CTC-236F12.4; RP11-155L15.1; IL6ST; AC008914.1; ANKRD55	MAP3K1; CTC-236F12.4; RP11-155L15.1; C5orf67; IL6ST; SLC38A9; CTD-2227I18.1	MAP3K1; AC008914.1; IL6ST; CDC20B; DDX4; ANKRD55; C5orf67; RP11-155L15.1
chr5:58388234-59569743	PDE4D	PDE4D	SETD9
chr5:90993653-91993653	RP11-213H15.1; LUCAT1; RP11-213H15.4; RP11-414H23.2	RP11-213H15.1; LUCAT1; RP11-213H15.4	ADGRV1; FAM81B; MCTP1; CTC-529L17.2; NR2F1; POU5F2; RP11-185E12.2; KIAA0825; SLF1; FAM172A; RP11-133F8.2; CTC-458G6.2
chr5:132571366-133571367	C5orf15; CTB-113I20.2; VDAC1; TCF7	C5orf15; TCF7; CTB-113I20.2; VDAC1	C5orf15; TCF7; WSPAR; CTB-113I20.2; VDAC1
chr5:158303005-159317075	CTC-436K13.1; CTC-436K13.5; EBF1; RP11-175K6.1	EBF1; RP11-175K6.1	
chr5:169664483-170664483			LINC01187; WWC1; PANK3
chr6:15899326-16899326	STMND1; ATXN1		STMND1

chr6:20121007-21121007	E2F3; MBOAT1; SOX4; RP11-204E9.1	E2F3; SOX4; DCDC2; CASC15; NBAT1; LINC00581; RP1-135L22.1	CASC15; NBAT1; DCDC2; MRS2; SOX4; KAAG1; RP11-204E9.1; E2F3; RP11-524C21.2; NRSN1
chr6:80918669-82086234	FAM46A; AL133475.1; TTK	FAM46A; BCKDHB	BCKDHB; ELOVL4; RP11-250B2.5
chr6:129527974-130527974	RP1-69D17.4; EPB41L2; ARHGAP18; TMEM244; AKAP7; RP11-102N11.1	ARHGAP18; RP1-69D17.4; TMEM244; RP11-102N11.1; EPB41L2	EPB41L2
chr6:151097720-152615881	ESR1; ARMT1; RMND1	ESR1; ARMT1; RMND1	ESR1
chr7:93984487-94984487	CASD1; PPP1R9A; PON1; AC002429.5; ASB4; AC004012.1	PEG10; SGCE; CASD1; ASB4; AC002429.5; AC004012.1; PON2; PPP1R9A; PON3	ASB4; AC002429.5; PON2; PON1; AC004012.1; PON3
chr8:60001-720692		LINC00115; LINC01128; RP5-857K21.4	MYOM2; KBTBD11; DLGAP2; RP11-439C15.4; RP11-439C15.5
chr8:29152099-30152100	DUSP4	DUSP4	
chr8:36500965-37501668	ERLIN2		
chr8:74818066-76005702			RP11-91P17.1; RP11-38H17.1
chr8:126412414-129029685	CASC11; MYC; CCDC26; RP11-419K12.1; PVT1; LINC00977; LINC00976; LINC00824	MYC; CASC11; CCDC26; RP11-419K12.1; PVT1; LINC00977; LINC00824; LINC00976	CCDC26; RP11-419K12.1; PCAT1; CASC11; MYC; RP11-26E5.1; PVT1; LINC00824; LINC00976; LINC00977
chr9:107041527-108633073	KLF4	KLF4	
chr10:8546150-9546150			RP5-1031D4.2; PRKCQ-AS1; RP11-554I8.1; SFMBT2; TAF3; RP5-1119O21.2; ITIH2
chr10:21244013-22620463	BMI1	MLLT10; RP11-573G6.10; BMI1; COMMD3; COMMD3-BMI1	MLLT10; SKIDA1
chr10:78581391-79627965	LINC00856; RPS24; RP11-90J7.3; POLR3A	RPS24; RP11-90J7.3; LINC00856; RP11-157J13.1	ZCCHC24; LINC00856; RP11-90J7.3; RP11-157J13.1; POLR3A; RPS24
chr10:112514168-113526395	TCF7L2; VTI1A; ZDHHC6; HABP2	TCF7L2; VTI1A; ZDHHC6; HABP2	
chr11:1377434-2421345	TNNT3	KRTAP5-6	
chr11:65276356-66315595	SIPA1	KAT5	
chr11:69509114-69521223*		CCND1	
chr11:107974789-108986410		DDX10	
chr12:13760997-14760997	ATF7IP	ATF7IP	
chr12:27486913-28881482	PTHLH	PTHLH	
chr12:114898717-115898717	TBX3; RBM19; RP11-162N7.1	TBX3; RP11-110L15.1; RP11-162N7.1	TBX3; RBM19; RP11-162N7.1; RP4-601P9.2; RP4-601P9.1; RP11-100F15.2; C12orf49; RP11-100F15.1; RP11-411G2.2; RPH3A; LHX5; RP11-25E2.1
chr12:119894342-120894343	PXN		
chr13:31894673-32898488		FRY	
chr14:36163563-37166547	SLC25A21	SLC25A21; MIPOL1	SLC25A21
chr14:67650477-69067965	ZFP36L1	ZFP36L1	

chr14:90874725-91902279		GPR68	
chr14:92137728-93150006			CTD-2547L24.3; GPR68; SERPINA5
chr15:90465983-91465985			FAM174B; CHD2
chr16:3556787-4556787	ADCY9; GLIS2-AS1	LINC01569	
chr16:52004913-53004913			CES5A
chr16:53267042-54321379	CRNDE; IRX5; IRX3; CTD-3032H12.2; CTD-3032H12.1	CRNDE; IRX5; IRX3; FTO; CTD-3032H12.2; MMP2; CTD-3032H12.1; RPGRIP1L; AC007491.1; RP11-434E6.4; RP11-324D17.1; RP11-324D17.2	CTD-3032H12.1; CTD-3032H12.2; CRNDE; IRX5; RP11-212I21.5; RP11-212I21.2; MMP2; ADCY7; RP11-26L20.4; AC007491.1; RP11-324D17.1; SLC6A2; RP11-212I21.4; IRX6; RP11-26L20.5; IRX3; LPCAT2; CTD-2302M15.1; NUDT21; OGFOD1; GNAO1; RP11-434E6.4; RP11-434E6.2
chr16:54148152-55148152		IRX3	
chr16:80114430-81117200			VAT1L
chr16:86551631-87551631	FOXC2; FOXL1; MTHFSD; RP11-58A18.1	FOXC2; FOXL1; MTHFSD	FOXL1; FOXC2; MTHFSD; RP11-442O1.3
chr19:12547463-14343759	IER2; STX10; LYL1; NACC1; TRMT1	LYL1; NFIX; NACC1; TRMT1; IER2; STX10	
chr19:16684212-17783315	HAUS8; MYO9B	HAUS8; MYO9B	
chr19:17939625-18960332	HOMER3; CERS1; GDF1; COPE; DDX49; C19orf60; UBA52	UBA52; C19orf60	
chr19:29286822-30286822			CTC-565M22.1; PDCD5; ZNF507; ANKRD27; RGS9BP; CTD-2538C1.2; NUDT19; CEP89
chr21:14701662-15701664		HSPA13	
chr22:37672826-39463350	PLA2G6; MAFF		
chr22:39980230-41131866	MKL1	MKL1	MKL1

Table 3.7: Distribution of GS2 'direct' genes at 129 breast cancer risk regions. Out of 129 fine-mapping regions to which 196 'strong-evidence' breast cancer risk signals map, there were 54 at which 145 unique genes formed direct IPs in the GS2 Dovetail 2kb dataset, 56 at which 156 unique genes formed direct IPs in the GS2 Arima 2kb dataset and 42 at which 177 unique genes formed direct IPs in the GS2 Standard dataset. Regions that were not covered by the capture array as well as regions where no putative target genes were identified using any of the three datasets are not shown. Fine-mapping region coordinates are in GRCh38/hg38. (*) – there were several fine-mapping regions (originally defined in hg19) that when lifted over to hg38 were split or partially deleted in hg38. These regions were compiled manually to encompass all CCVs at each of the regions. Purple – genes that were called on the basis of *trans* interaction peaks.

At most regions, the Arima and Dovetail protocols identified 1 – 3 putative target genes, while the Standard protocol identified 4 or more genes at more than 40% of the regions, making it less feasible to perform functional follow up on these genes.

Protocol	T-47D		GS2	
	≤ 3 genes	≥ 4 genes	≤ 3 genes	≥ 4 genes
Dovetail	33 (82.5%)	7 (17.5%)	37 (68.5%)	17 (31.5%)
Arima	30 (85.7%)	5 (14.3%)	41 (73.2%)	15 (26.8%)
Standard	23 (54.8%)	19 (45.2%)	25 (59.5%)	17 (40.5%)

Table 3.8: Number of putative target genes per region. Absolute numbers and percentages (in brackets) of breast cancer risk fine-mapping regions where ≤ 3 or ≥ 4 putative target genes were identified per region using the 2kb Dovetail, 2kb Arima and Standard protocol rCHi-C data in T-47D and GS2 cells.

Focusing next on 'direct' CCVs identified in the 2kb Dovetail, 2kb Arima and Standard datasets, these were mapped back to 196 'strong-evidence' breast cancer risk signals. This resulted in at least one CCV identified at 77 signals in at least one T-47D dataset, and at 93 signals in at least one GS2 dataset. These 106 signals together with the number of CCVs that formed direct interaction peaks in these six datasets are shown in Table 3.9. There was one additional signal (signal 2 at the chr12:27,486,913-28,881,482 region) which was not targeted by the rCHi-C capture array, but some of its CCVs (that mapped by chance to the non-baited interacting bin) formed direct interaction peaks with the *PTHLH* promoter (which maps to the baited region).

Fine-mapping region	Signal	Index SNP	Number of CCVs	Captured CCVs	T-47D Dovetail 2kb	T-47D Arima 2kb	T-47D Standard	GS2 Dovetail 2kb	GS2 Arima 2kb	GS2 Standard
chr1:9983762-11006158	Signal 1	rs657244	19	1	0	0	1	1	0	1
chr1:45635245-46635245	Signal 1	rs12039667	11	7	1	0	0	2	0	0
chr1:87191240-88191240	Signal 2	rs11583393	8	8	0	1	6	8	8	8
chr1:113405767-114405767	Signal 1	rs11102701	12	12	1	4	0	6	6	0
chr1:145625092-146010623*	Signal 1	rs143384623	35	35	0	2	0	1	2	0
chr1:145625092-146010623*	Signal 2	rs200366104	5	1	1	1	0	1	1	0
chr1:154676305-155678990	Signal 1	rs1057941	16	16	7	0	5	3	1	4
chr1:203332121-204332121	Signal 1	rs59867004	56	56	9	7	11	6	7	10
chr1:204049714-205049714	Signal 1	rs4951401	6	6	2	1	4	0	1	0
chr1:216547232-217547232	Signal 1	rs11117754	12	11	1	0	0	0	0	0
chr2:18634525-19621042	Signal 1	rs10184522	17	17	0	0	0	0	5	3
chr2:28447809-29447810	Signal 1	rs71403627	81	81	0	0	17	1	4	9
chr2:119987546-120988992	Signal 2	rs4849879	16	16	0	0	5	0	0	0
chr2:119987546-120988992	Signal 4	rs17625845	6	5	0	0	6	0	1	4
chr2:171019711-172608243	Signal 1	rs2016394	3	3	0	0	0	2	3	3
chr2:171019711-172608243	Signal 2	rs13020413	35	35	0	1	6	5	29	30
chr2:172846180-173848166	Signal 1	rs7589172	13	13	0	0	0	2	0	1
chr2:200816524-201816524	Signal 2	rs13015648	8	6	0	0	0	3	2	0
chr2:216541109-217931785	Signal 1	rs4442975	1	1	1	0	0	1	0	1
chr2:216541109-217931785	Signal 2	rs138522813	5	5	2	0	0	2	2	0
chr2:216541109-217931785	Signal 3	rs5838651	42	42	0	0	30	19	2	42
chr3:4200592-5200591	Signal 1	rs6787391	4	4	4	4	1	1	3	3
chr3:26786474-28243756	Signal 1	rs1352944	44	44	0	5	23	1	11	18
chr3:26786474-28243756	Signal 2	rs36078735	12	12	0	0	5	0	1	7
chr3:63456021-64482224	Signal 1	rs555060306	94	94	3	7	14	12	11	3
chr3:86488393-87488393	Signal 1	rs13066793	2	2	0	1	0	1	2	0
chr3:140894017-141894017	Signal 1	rs7625643	24	24	4	0	23	12	0	0
chr4:82948971-83948971	Signal 1	rs6854739	84	84	0	0	0	6	7	0
chr4:104647856-105935604	Signal 1	rs17617028	21	20	5	2	1	4	6	4
chr5:779675-1797374	Signal 1	rs10069690	1	1	0	0	1	0	0	0
chr5:779675-1797374	Signal 2	rs2736107	5	5	0	0	5	0	0	0
chr5:779675-1797374	Signal 3	rs150804576	23	7	0	0	6	0	0	0
chr5:15687249-16687419	Signal 1	rs12652713	45	45	0	0	7	0	0	0
chr5:44013202-45206396	Signal 3	rs13153426	72	65	2	1	3	7	1	1
chr5:56236057-57292056	Signal 3	rs112497245	21	21	10	12	15	6	8	10
chr5:56236057-57292056	Signal 4	rs7730210	70	38	0	15	6	4	24	6
chr5:58388234-59569743	Signal 1	rs537267133	30	30	0	0	0	0	0	1

chr5:58388234-59569743	Signal 2	rs10472097	5	1	1	0	0	2	2	0
chr5:90993653-91993653	Signal 1	rs1964292	88	88	18	15	30	9	11	12
chr5:132571366-133571367	Signal 1	rs571173399	117	117	0	0	17	12	5	21
chr5:158303005-159317075	Signal 1	rs31864	5	5	2	0	1	4	1	0
chr5:169664483-170664483	Signal 1	rs56722914	19	19	0	0	0	0	0	4
chr6:15899326-16899326	Signal 1	rs3819405	1	1	0	0	0	1	0	1
chr6:20121007-21121007	Signal 1	rs2328531	52	36	2	2	1	2	2	12
chr6:80918669-82086234	Signal 1	rs7763102	51	41	0	0	4	5	8	5
chr6:129527974-130527974	Signal 1	rs6569648	43	43	2	0	0	12	5	20
chr6:151097720-152615881	Signal 2	rs34133739	1	1	0	0	0	1	1	0
chr6:151097720-152615881	Signal 5	rs79388591	173	173	0	0	0	5	0	8
chr6:151097720-152615881	Signal 6	rs9918437	22	6	0	1	0	0	0	0
chr7:93984487-94984487	Signal 1	rs1879854	47	42	2	2	10	9	23	25
chr8:60001-720692	Signal 1	rs34810249	25	24	0	0	0	0	6	1
chr8:29152099-30152100	Signal 1	rs7465364	16	16	0	0	0	6	1	0
chr8:36500965-37501668	Signal 1	rs4286946	16	16	1	0	0	1	0	0
chr8:74818066-76005702	Signal 3	rs17303163	16	16	0	0	4	0	0	4
chr8:100966731-101966731	Signal 1	rs7813150	47	47	16	0	0	0	0	0
chr8:123097926-124097925	Signal 1	rs4871411	23	23	0	1	0	0	0	0
chr8:126412414-129029685	Signal 1	rs10096351	3	3	0	0	2	0	0	2
chr8:126412414-129029685	Signal 2	rs7017073	44	44	14	16	10	39	41	40
chr8:126412414-129029685	Signal 3	rs35961416	1	1	0	0	1	0	0	1
chr8:126412414-129029685	Signal 4	rs419018	43	43	0	0	39	0	0	38
chr9:107041527-108633073	Signal 1	rs659713	10	10	0	0	0	3	2	0
chr9:107041527-108633073	Signal 3	rs10816625	1	1	0	0	0	0	1	0
chr9:107041527-108633073	Signal 4	rs13294895	1	1	0	0	0	0	1	0
chr10:8546150-9546150	Signal 1	rs7081544	49	48	44	19	39	0	0	8
chr10:21244013-22620463	Signal 1	rs7098100	7	7	0	0	0	1	2	0
chr10:21244013-22620463	Signal 2	rs138026227	58	16	0	0	0	0	1	1
chr10:78581391-79627965	Signal 1	rs754416	10	10	0	0	9	3	4	9
chr10:78581391-79627965	Signal 2	rs10762851	17	17	4	0	17	4	0	17
chr10:78581391-79627965	Signal 3	rs61862474	9	9	0	0	8	0	0	8
chr10:112514168-113526395	Signal 1	rs12250948	12	12	0	0	0	5	3	0
chr11:1377434-2421345	Signal 1	rs620315	7	7	2	0	7	4	2	0
chr11:65276356-66315595	Signal 1	rs548082010	13	12	3	0	1	3	3	0
chr11:69509114-69521223*	Signal 3	rs671888	20	20	0	0	0	0	3	0
chr11:107974789-108986410	Signal 1	rs368848598	59	25	8	2	1	0	1	0
chr11:129082612-130091276	Signal 1	rs745382	17	17	1	7	6	0	0	0
chr12:13760997-14760997	Signal 1	rs12422552	18	18	9	7	0	11	9	0

chr12:27486913-28881482	Signal 2	rs1600346	375	0	0	9	0	21	51	0
chr12:114898717-115898717	Signal 1	rs1353783	7	7	0	2	7	0	2	0
chr12:114898717-115898717	Signal 2	rs35422	1	1	0	0	1	0	0	1
chr12:114898717-115898717	Signal 3	rs1882155	8	8	0	0	8	5	7	3
chr12:114898717-115898717	Signal 4	rs11067765	6	6	5	5	6	4	4	6
chr12:119894342-120894343	Signal 1	rs184486140	5	5	3	3	1	1	0	0
chr13:31894673-32898488	Signal 1	rs11571833	5	4	0	0	0	0	1	0
chr14:36163563-37166547	Signal 1	rs12881240	19	19	3	7	2	5	12	2
chr14:36163563-37166547	Signal 2	rs848088	20	20	2	0	0	0	0	0
chr14:67650477-69067965	Signal 1	rs35378451	8	8	1	0	0	8	3	0
chr14:67650477-69067965	Signal 2	rs2478777	4	4	3	0	1	3	1	0
chr14:90874725-91902279	Signal 1	rs11341843	3	3	0	0	0	0	1	0
chr14:92137728-93150006	Signal 1	rs78440108	34	34	0	0	0	0	0	1
chr15:90465983-91465985	Signal 1	rs12594752	22	22	0	1	6	0	0	6
chr16:3556787-4556787	Signal 1	rs6500580	23	14	0	1	0	1	3	0
chr16:3556787-4556787	Signal 2	rs8063564	14	12	0	1	0	1	1	0
chr16:52004913-53004913	Signal 1	rs4784227	1	1	0	0	0	0	0	1
chr16:53267042-54321379	Signal 1	rs55872725	6	6	6	5	6	6	6	6
chr16:53267042-54321379	Signal 2	rs9925952	21	21	0	0	0	16	15	16
chr16:54148152-55148152	Signal 1	rs28539243	3	3	0	0	0	0	2	0
chr16:80114430-81117200	Signal 1	rs9938021	14	14	3	1	14	0	0	5
chr16:86551631-87551631	Signal 1	rs4066743	85	85	10	28	21	28	55	21
chr17:79794855-79816335*	Signal 1	rs2587505	10	10	6	7	7	0	0	0
chr19:12547463-14343759	Signal 1	rs78269692	21	8	1	1	0	2	2	0
chr19:16684212-17783315	Signal 1	rs67397200	16	16	0	0	0	1	2	0
chr19:17939625-18960332	Signal 1	rs8105994	56	56	0	0	0	13	4	0
chr19:29286822-30286822	Signal 1	rs17513613	60	47	0	0	7	0	0	29
chr21:14701662-15701664	Signal 1	rs2403907	7	7	1	0	0	0	0	0
chr21:14701662-15701664	Signal 2	rs2822999	29	29	0	3	0	0	2	0
chr22:37672826-39463350	Signal 1	rs5995543	27	27	0	0	0	3	0	0
chr22:39980230-41131866	Signal 1	rs66987842	196	196	0	0	0	36	14	1

Table 3.9: Distribution of numbers of 'direct' CCVs identified using T-47D and GS2 rCHi-C data by breast cancer risk signals. 106 signals at which my capture array covered at least one CCV and at which at least one CCV formed direct IPs in at least one out of six rCHi-C datasets. Number of CCVs – number of CCVs reported for that signal; Captured CCVs – number of CCVs that were targeted by the capture array. Fine-mapping region coordinates are in GRCh38/hg38. (*) – there were several fine-mapping regions (originally defined in hg19) that when lifted over to hg38 were split or partially deleted in hg38. These regions were compiled manually to encompass all CCVs at each of the regions. Red – a signal that was not targeted by the rCHi-C capture array, but at which some of the CCVs (that mapped 'by chance' to the non-baited interacting bins) were involved in direct IPs.

To estimate which protocol performs better in narrowing down the number of CCVs, I compared numbers across the protocols (Table 3.10). This revealed that the Dovetail and Arima datasets tend to prioritise fewer CCVs than the Standard protocol.

	Arima > Standard	Arima < Standard	Arima = Standard
T-47D	10	16	1
GS2	11	17	6
	Dovetail > Standard	**Dovetail < Standard**	**Dovetail = Standard**
T-47D	12	15	1
GS2	7	20	8
	Arima > Dovetail	**Arima < Dovetail**	**Arima = Dovetail**
T-47D	8	10	7
GS2	24	18	12

Table 3.10: Comparison of numbers of 'direct' CCVs identified per signal. Comparison of numbers of breast cancer risk signals at which each of the protocols identified more, less or the same number of 'direct' CCVs per signal. 2kb Dovetail, 2kb Arima and Standard protocol rCHi-C data in T-47D and GS2 cells were used.

Based on all the above comparisons, the Arima and Dovetail protocols seemed to perform at least as well or better than our Standard in-house protocol, in addition to being more cost- and time-effective, requiring less cellular input and resulting in increased resolution. Therefore, it seemed reasonable to use one of the kit-based methods for CHi-C library generation in primary luminal epithelial and fibroblast cells. For pragmatic reasons related to the COVID-19 pandemic and the high level of technical support from the company, I selected the Dovetail Genomics protocol as the protocol of choice.

3.7. Discussion

Capture Hi-C is a chromosome conformation capture-based method that allows high-throughput and high-resolution analysis of physical interactions between regulatory elements and their target genes. Although multiple CHi-C protocols have been developed and optimised over the years, no gold-standard method is available yet, with each of the protocols having their own advantages and disadvantages. The choice of the most suitable protocol therefore depends on the project aims.

Here I compared three methods: our Standard in-house protocol, the Arima Genomics Hi-C kit and the Dovetail Genomics Omni-C kit (Table 3.11). Major limitations of the Standard (HindIII) protocol are the resolution (~ 10 kb), the non-random distribution of

restriction sites and the requirement for a very high cellular input. Using kit-based methods, I managed to increase the resolution to ~ 2 kb, using far lower cellular inputs. In addition, the Arima and Dovetail protocols were proven to be less time-consuming and more cost-effective.

	Standard in-house protocol	**Arima Hi-C kit**	**Dovetail Genomics Omni-C kit**
Cellular input	6 x 10^7 cells	2 x 10^6 cells	1 x 10^6 cells
Total time (days)	15.5 days	7 days	6 days
Crosslinking time	1 day	1 hour	1 hour
Hi-C library generation time	8.5 days	4 days	3 days
Target enrichment and post-hybridisation amplification	3 days	3 days	3 days
Quality Control(s)	3 days (separate)	3h (integrated into the protocol)	3h (integrated into the protocol)
Estimated cost	£2120	£850	£660
Resolution	~ 10 kb	2 kb	2 kb
Dangling end removal step	YES	NO	NO

Table 3.11: Comparison of the standard in-house, the Arima Hi-C and the Dovetail Genomics Omni-C protocols. Cellular input, time and estimated costs were calculated based on one complete rCHi-C library generated from GS2 cells. Time and costs associated with cell culturing and next-generation sequencing are not included. The materials required for the quality controls are included for the Arima Hi-C and Dovetail Genomics Omni-C protocols, but are not included for the Standard protocol (since its QCs are not integrated into the protocol).

However, the Standard libraries had the smallest proportion of very short-range interactions (*cis* < 1 kb) that often represent dangling ends due to partial digestion of chromatin (< 1% compared to approximately 10% and 20% in the Arima and Dovetail libraries). This is in part because the average size of the fragments in Standard libraries limits the potential for such interactions, but also because the protocol includes a step for removing 'dangling ends'. In addition, 3C-based methods are not considered reliable for detecting interaction peaks over distances of less than 10 kb[115], therefore, we cannot assess whether the read pairs where the distance between interacting fragments (when mapped back to the reference genome) is less than 10 kb represent 'true' interactions and

we have to consider them as uninformative. As a result, the Dovetail and Arima libraries might require more sequencing to get a similar amount of 'usable' data; however, the proportions of unique and on-target read pairs would probably have a greater impact on the overall amount of 'usable' data.

Interaction calling using CHiCANE[108] revealed that Standard libraries had the lowest proportions of *cis* interaction peaks that were called within the 100 kb – 1 Mb range (32% and 46% compared to > 80% in the 2kb-binned Arima and Dovetail libraries) and the highest proportion of interaction peaks called in the ≥ 1 Mb range (~ 50% compared to < 10% in the 2kb-binned Arima and Dovetail libraries). As the majority of regulatory interactions that drive GWAS associations are likely to be within the 100 kb to 1 Mb range[110, 111], this favours the Arima and Dovetail protocols.

To identify and prioritise causal variants and target genes that may be involved in mediating breast cancer risk associations, I mapped CHi-C data to annotated gene promoters and breast cancer risk CCVs, and specifically looked into the subset of interaction peaks in which a bin colocalising with a gene promoter formed a direct interaction peak with a CCV-containing bin.

Comparing across the protocols, I observed that, on average, 'direct' genes and CCVs identified in the Dovetail libraries had higher overlap with those in the Arima libraries rather than with those in the Standard libraries (Figure 3.2 and Figure 3.3). In addition, there was higher overlap between the Arima and Standard libraries, than between the Dovetail and Standard libraries. This may reflect the fact that the Arima and Standard protocols are both RE-based and have higher protocol similarity, or this could be explained by differences in downstream analysis. Specifically, the Standard protocol libraries were called using individual HindIII fragments as the unit of analysis, the Arima libraries were called using fixed size bins (2kb and 5kb), while the Dovetail libraries were called using 2kb- and 5kb-binned data with the gene promoters allocated into the smart bins that can vary in lengths (median = 1.9 kb). See Section 6.1 for more details.

Generally, the largest numbers of non-replicated genes and CCVs (i.e., 'direct' genes and CCVs that were called in a single dataset) were found in libraries generated using the Standard protocol (Figure 3.2 and Figure 3.3). Analysis revealed that half (T-47D) and one-third (GS2) of 'direct' genes that were only identified in the Standard protocol

libraries mapped to very long-range or *trans* interaction peaks (Table 3.5). As mentioned previously, the available evidence suggests that these are less likely to represent functional enhancer-promoter interaction peaks than those that occur within a distance of approximately 100 kb – 1 Mb[110, 111, 116]. Although some very long-range or even *trans* interaction peaks might represent valid interacting events, the fact that they appear to be method specific rather suggests that they are more likely to represent chance findings. To further address this issue, it would be informative to compare whether such interaction peaks are replicated across multiple technical replicates. Very long-range and *trans* interaction peaks also explain a (lower) proportion of the non-replicated CCVs that were identified in the Standard libraries. In this instance, however, at least a subset of the remaining non-replicated CCVs may be the result of the non-random distribution of restriction sites leading to some very large HindIII fragments that contain many CCVs.

Since the true number of causal variants and target genes underlying each association signal is unknown, it is difficult to assess the quality of data generated by different protocols by examining the numbers alone. Therefore, I decided to investigate the 2q35 breast cancer risk locus as a 'proof of principle' locus that has been extensively functionally characterised[101, 112-114] and to check whether my rCH-C data would select direct interaction peaks between rs4442975 and esv3594306 (the likely causal variants at signals 1 and 2, respectively) and promoter of the likely target gene *IGFBP5*. Out of 10 datasets, 4 picked up a direct interaction peak between the *IGFBP5* promoter and rs4442975 (Standard GS2, 5kb Arima T-47D, 2kb Dovetail GS2 and T-47D) and 3 picked up a direct interaction peak between the *IGFBP5* promoter and esv3594306 (2kb Dovetail GS2 and T-47D, 2kb Arima GS2).

Interestingly, more activity was observed at the 2q35 locus in GS2 libraries rather than in T-47D libraries. RNA-seq and H3K27ac CUT&Tag data generated by other members of the Functional Genetic Epidemiology lab showed that *IGFBP5* expression is significantly higher in GS2 cells than in T-47D cells (2.92-fold; $p = 1.2 \times 10^{-8}$), and that there are marginally more GS2-specific (N=10) than T-47D-specific (N=8) H3K27ac peaks. However, rs4442975 and esv3594306 both colocalise with a T-47D-specific peak, not a GS2-specific peak (Figure 3.4 and Figure 3.5), suggesting that signals 1 and 2 are likely to mediate the association with breast cancer risk via epithelial cells rather than fibroblasts.

Comparing the ability of each protocol to prioritise a subset of putative target gene and CCVs, my data suggest that the Dovetail and Arima libraries tend to narrow down the numbers of genes and CCVs better than the Standard libraries (Table 3.8 and Table 3.10). Although it is possible that some of the genes and/or CCVs that were identified in the Standard libraries alone represent 'true' target genes and functional variants that were simply missed by the Dovetail and Arima protocols, the goal of my study was to prioritise putative target genes and CCVs that warrant in-depth functional follow up, and my overall conclusion is that it would be easier to do this using one of the kit-based methods rather than the Standard protocol.

Thus, based on the available information, the kit-based methods seemed to perform at least as well or better than our Standard in-house protocol, in addition to being more cost- and time-effective, requiring less cellular input and resulting in increased resolution. Therefore, it seemed reasonable to use one of the kit-based methods for CHi-C library generation in primary luminal epithelial and fibroblast cells.

4. Promoter Capture Hi-C in T-47D and GS2 cells

4.1. Overview of the libraries

Several studies have used pCHi-C to characterise GWAS risk loci[117-120]. To compare rCHi-C and pCHi-C approaches, I generated pCHi-C libraries in T-47D and GS2 cell lines using the Dovetail Genomics Omni-C protocol combined with their promoter enrichment panel. The Dovetail Pan Promoter Enrichment Panel targets over 98% of human promoter regions (84,643 promoters associated with 27,375 coding and non-coding genes) and, for the purposes of downstream analysis, allocates these to 39,825 smart bins. Most smart bins capture promoter(s) of a single gene, but in some smart bins (N=2,905) the proximity of gene promoters is such that a single smart bin captures multiple promoters. In addition, a subset of genes (N=9,104) occupies more than one smart bin.

Significant interaction peaks were called in the 2kb- and 5kb-binned data using CHiCANE. The number of interaction peaks varied from 22,401 to 48,863 (Table 4.1). As in the rCHi-C libraries, a much larger number of interaction peaks were called in the GS2 dataset than in the T-47D dataset. Comparing across the bin sizes, the 2kb-binned datasets showed lower proportions of *cis* interaction peaks in the 100 kb – 1 Mb range and higher proportions of *cis* interaction peaks in the 10 kb – 100 kb range. Comparing back to the rCHi-C libraries (Table 3.2), the overall proportions of *trans* and *cis* ≥ 1 Mb interaction peaks were generally lower in pCHi-C than in rCHi-C libraries. In the 5kb-binned rCHi-C libraries, the proportions of *cis* interaction peaks in the ≥ 1 Mb range were over 50%, while in pCHi-C libraries only about 4% (T-47D) and 7% (GS2) of interaction peaks were in this range.

4.2. Direct interaction peaks

To compare pCHi-C to a rCHi-C approach, I mapped the pCHi-C data to the 5,117 CCVs associated with 196 ‘strong-evidence’ breast cancer risk GWAS signals and selected direct interaction peaks (Table 4.2). The number of direct interaction peaks varied from 79 to 248. They involved 36 to 105 unique smart bins harbouring a total of 43 to 107 ‘direct’ genes (1 to 3 genes per bin).

Cell line	Bin size	Total IPs	*trans* IPs		*cis* < 1kb		*cis* 1kb - 10kb		*cis* 10kb - 100kb		*cis* 100kb - 1Mb		*cis* ≥ 1Mb	
T-47D	5kb	28,157	147	0.52%	0	0%	29	0.1%	2,611	9.3%	24,125	86%	1,245	4%
T-47D	2kb	22,401	78	0.35%	0	0%	48	0.2%	5,562	24.8%	16,318	73%	395	2%
GS2	5kb	48,863	53	0.11%	0	0%	11	0.02%	3,655	7.5%	41,956	86%	3,188	7%
GS2	2kb	48,529	42	0.09%	0	0%	15	0.03%	8,966	18.5%	38,594	80%	912	2%

Table 4.1: Interaction peak calling statistics for T-47D and GS2 pCHi-C libraries. A breakdown of significant (q-value ≤ 0.1) interaction peaks called in pCHi-C libraries generated in T-47D and GS2 cell lines using the Dovetail Genomics Omni-C protocol and called using CHiCANE. Libraries were called using both 2kb- and 5kb-binned data.

Cell line	Bin size	Total direct IPs	Unique gene-containing bins	Unique genes	Unique CCV-containing bins	Unique CCVs	CCVs per bin
T-47D	5kb	138	61	61	102	281	1 to 15
T-47D	2kb	79	36	43	70	130	1 to 8
GS2	5kb	248	105	107	172	528	1 to 19
GS2	2kb	147	72	80	130	223	1 to 10

Table 4.2: Summary of direct interaction peaks called in T-47D and GS2 pCHi-C libraries. Direct interaction peaks – interaction peaks in which a bin colocalising with a gene promoter forms a direct interaction with a CCV-containing bin.

Overall, 130 to 528 CCVs were involved in direct interaction peaks ('direct' CCVs). On average, larger numbers of CCVs per bin were observed in the 5kb datasets rather than in the 2 kb datasets (T-47D: 1 to 15 vs. 1 to 8; GS2: 1 to 19 vs. 1 to 10). The overall trend was different to that in the rCHi-C libraries, with larger numbers of interaction peaks, genes and CCVs identified in the 5kb datasets versus 2kb datasets. In order to compare the genes and CCVs that would be prioritised by the two approaches (pCHi-C and rCHi-C) directly and to look for the third-party interaction peaks (Section 4.3), I needed to use the same bin size for both types of data, so I focused on the 2kb-binned data for the rest of the analysis.

There were approximately twice as many direct interaction peaks as well as gene- and CCV-containing bins in the rCHi-C (compared to pCHi-C) datasets. As a result, many more non-replicated 'direct' genes (and CCVs) were observed in the rCHi-C datasets than in pCHi-C ones (Figure 4.1). In T-47D cells, 31 genes formed direct interaction peaks in both rCHi-C and pCHi-C datasets, 64 genes were involved in direct interaction peaks in the rCHi-C dataset exclusively, and 12 'direct' genes were uniquely identified in the pCHi-C dataset (GS2: in both – 64; rCHi-C only – 81; pCHi-C only – 16). Some of the genes that were 'unique' to the pCHi-C datasets, such as *GMIP*, *WNT7B*, *CTD-2203A3.1*, *LINC01556* and *GATAD1* (GS2 only), were targets of fine-mapping regions that were excluded from my rCHi-C capture array (Section 2.5), meaning that I could not pick these genes up using my rCHi-C data. Others (T-47D: 4 genes; GS2: 5 genes) mapped to the regions where there were no 'direct' genes in the rCHi-C data. Two such examples are *CCND1* and *MYEOV* which formed direct interaction peaks in the GS2 pCHi-C dataset at the chr11:69,509,114-69,521,223 fine-mapping region (11q13.3 locus; Figure 4.2). The *CCND1* promoter formed two direct interaction peaks – one with rs35039974 and one with a bin containing rs476679, rs602690 and rs510754. The *MYEOV* promoter also formed two direct interaction peaks, but with different CCVs – one with a bin containing rs573073, rs2510848 and rs491193, and the second one with a bin containing rs678214, rs493786, rs676856, rs640822, rs665095, rs2015489 and rs120206.

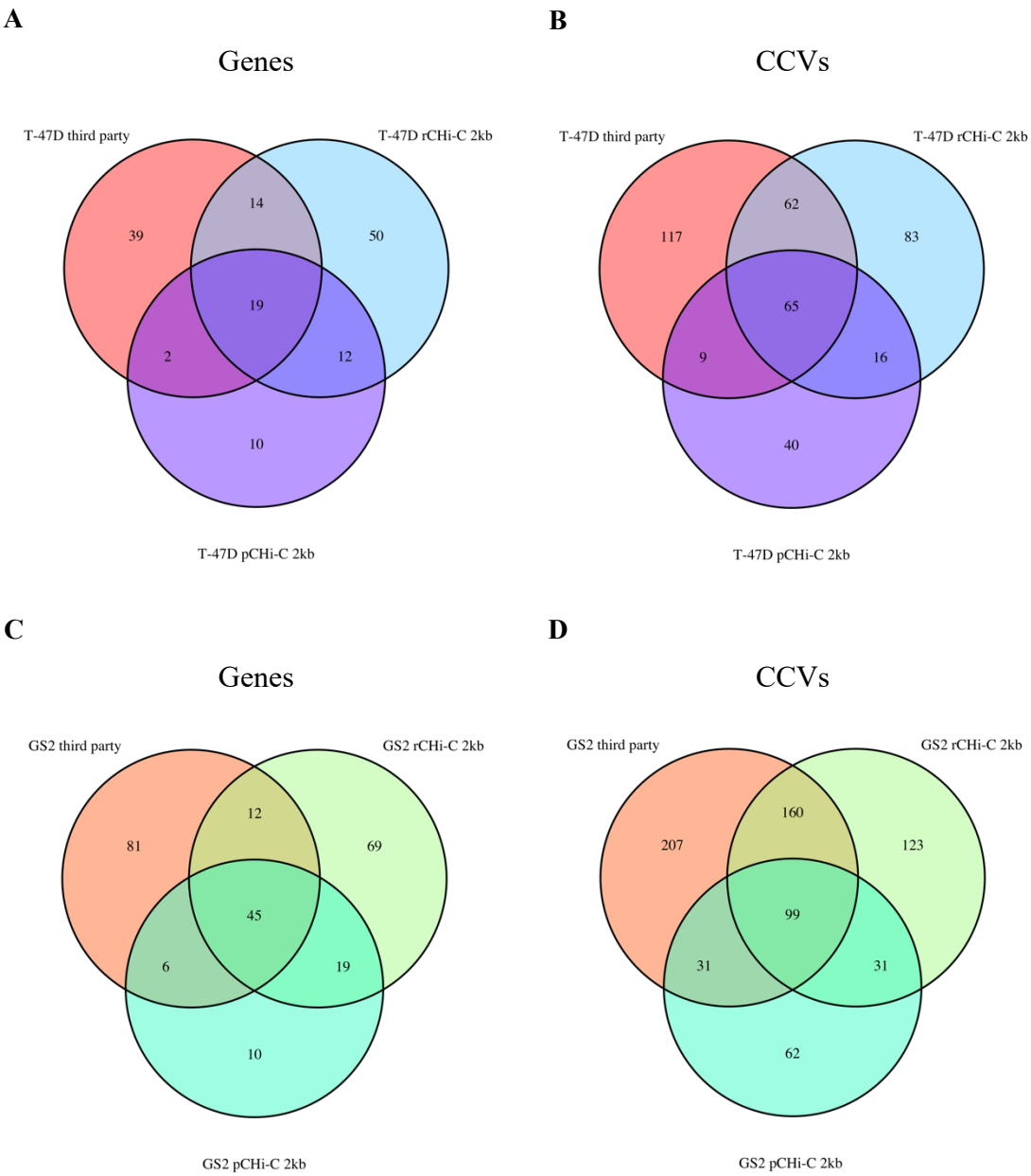

Figure 4.1: Venn diagrams illustrating the overlap between 'direct' and 'third-party' genes and CCVs. (A, C) The overlap between 'third-party' genes and 'direct' genes identified in the rCHi-C and pCHi-C datasets generated in T-47D and GS2 cell lines. (B, D) The overlap between 'third-party' CCVs and 'direct' CCVs identified in the rCHi-C and pCHi-C datasets. 2kb-binned rCHi-C and pCHi-C datasets generated using the Dovetail Genomics Omni-C protocol were used.

To investigate to what extent the 'direct' genes and CCVs that were replicated between rCHi-C and pCHi-C datasets were replicated on the basis of identical direct interaction peaks (as opposed to simply being involved in any direct interaction peak), I compared the direct interaction peaks identified by both methods (Figure 4.3). Out of 192 and 79 direct interaction peaks identified in T-47D rCHi-C and pCHi-C libraries, respectively, 45 were reciprocal and involved 28 genes and 80 CCVs. In GS2, there were 87 reciprocal

interaction peaks that involved 61 genes and 124 CCVs. Therefore, almost 60% of direct interaction peaks identified in pCHi-C were replicates of rCHi-C interaction peaks.

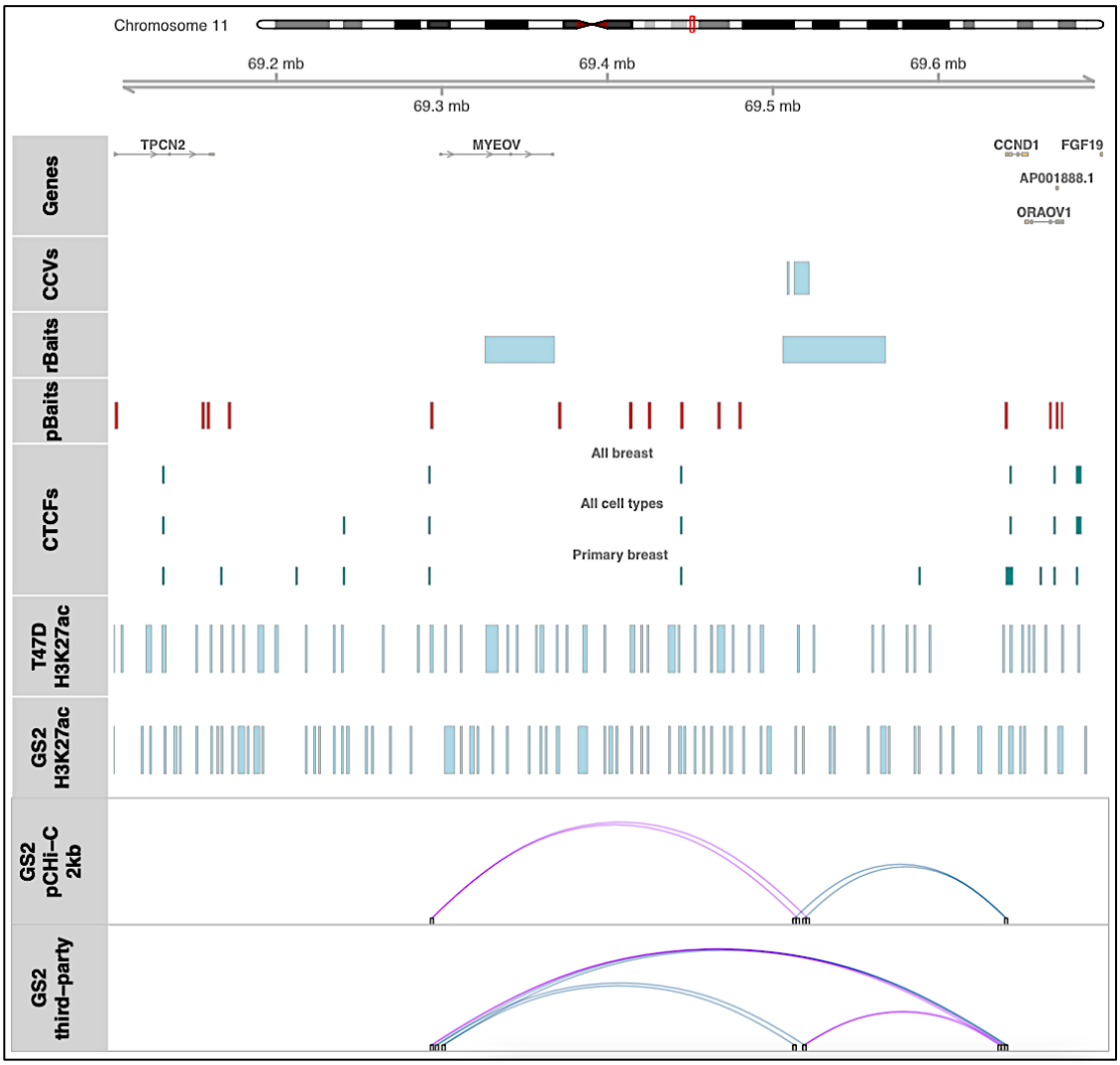

Figure 4.2: Interaction peaks at 11q13.3 involving *CCND1* and *MYEOV* promoters. Direct interaction peaks (from pCHi-C data) and third-party interaction peaks identified in GS2 cells at the 11q13.3 breast cancer risk locus (chr11:69,509,114-69,521,223 fine-mapping region, hg38) that involved *CCND1* (blue loops) and *MYEOV* (purple loops) promoters. Two other 'third-party' genes (*AP000439.2* and *RP11-554A11.7*) that also formed third-party interaction peaks at this locus are not shown. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. pBaits – pCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). H3K27ac – H3K27ac peaks identified from CUT&Tag data generated in T-47D and GS2 cell lines.

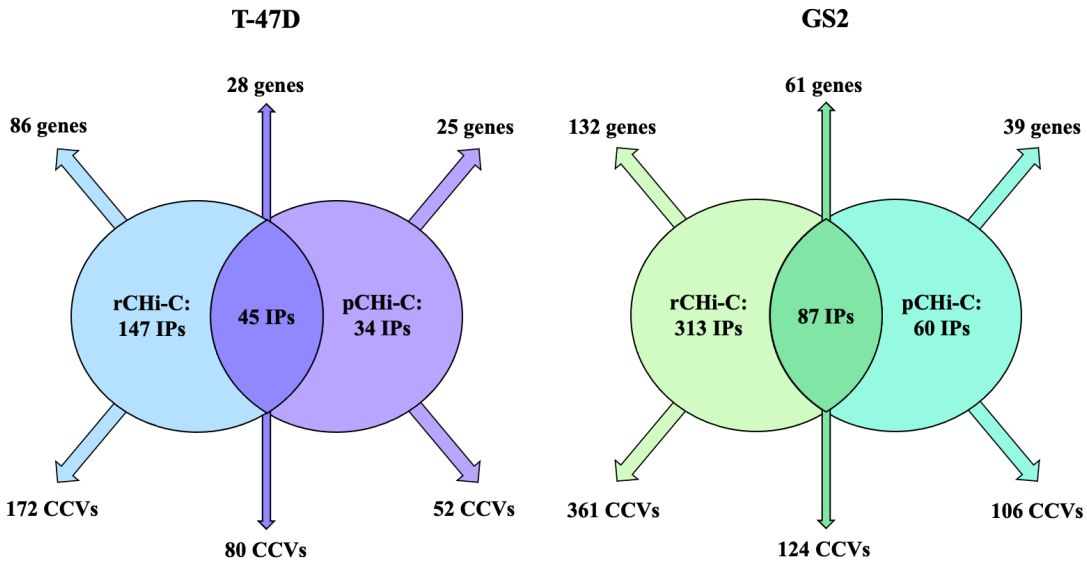

Figure 4.3: A breakdown of direct interaction peaks identified in rCHi-C and pCHi-C datasets. Out of 192 and 79 direct interaction peaks identified in T-47D rCHi-C and pCHi-C libraries, respectively, 45 were reciprocal and involved 28 genes and 80 CCVs. In GS2, there were 87 reciprocal interaction peaks that involved 61 genes and 124 CCVs. The sum of the genes indicated in this figure differs from the number of unique 'direct' genes reported in Table 3.4 (e.g., 86+28=114 vs. 95 for 2kb Dovetail T-47D rCHi-C dataset), as a gene can appear in multiple categories: (i) in the rCHi-C only category (blue); (ii) in the rCHi-C only (blue) and in the rCHi-C and pCHi-C overlap (middle, dark purple) categories; (iii) in the rCHi-C only (blue) and in the pCHi-C only (light purple) categories . This is because the same gene can form different direct interaction peaks, and the overlap in this figure is done on the basis of interaction peaks.

4.3. Third-party interaction peaks

So far, I have made the assumption that a target gene forms a direct interaction peak with a functional variant. However, a recent analysis of promoter-interacting expression quantitative trait loci defined a subset of 'indirect' interaction peaks[121]. To explore the possibility that a target gene and a CCV might be brought into proximity with each other by forming interaction peaks with a 'third party', I searched for non-baited target bins that formed both a statistically significant interaction peak with a baited smart bin in the pCHi-C data and with a baited CCV-containing bin in the rCHi-C data (hereafter, third-party bins). In T-47D cells, there were 174 third-party bins that were involved in 219 interaction peaks in the pCHi-C data and in 520 interaction peaks in the rCHi-C data, giving rise to 652 potential unique 'CCV – third party – gene' combinations (Table 4.3 and Figure 4.4). Overall, these third-party interaction peaks involved 74 genes and 253 CCVs. In GS2, 488 third-party bins formed 609 and 1,539 interaction peaks in the pCHi-C and rCHi-C datasets, respectively, resulting in 2,135 possible 'CCV – third party – gene' combinations that involved 144 genes and 497 CCVs (Table 4.3 and Figure 4.4).

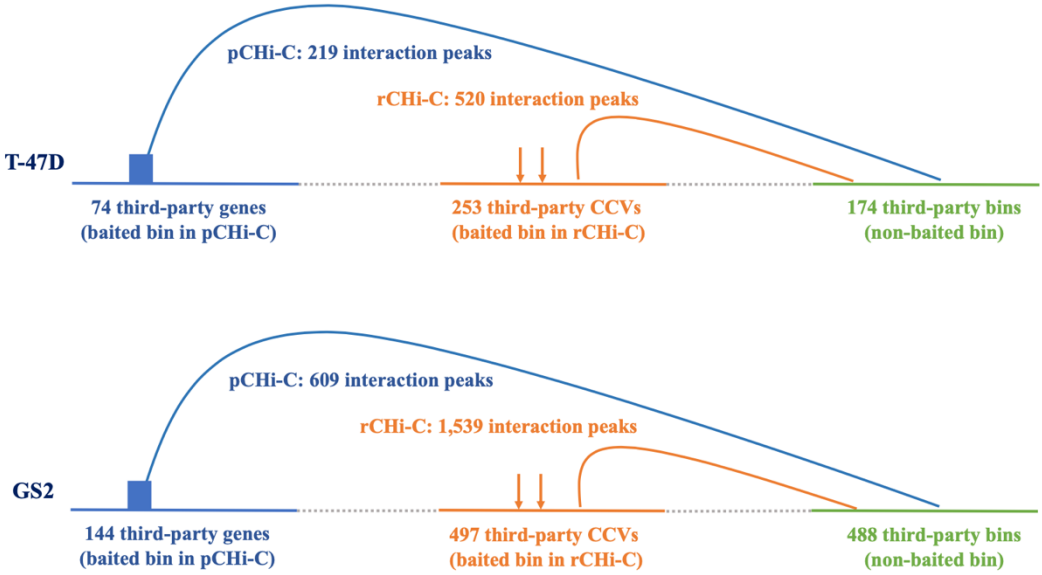

Figure 4.4: Summary of third-party interaction peaks called in T-47D and GS2 libraries. pCHi-C baited bins (blue) can form interaction peaks with a non-baited, non-CCV-containing target bin (green). If the same non-baited target bin formed interaction peaks with a baited CCV-containing bin (orange) in the rCHi-C data, then this non-baited target bin was considered to be a 'third-party' bin that mediated an 'indirect' interaction peak. Genes are indicated by blue rectangles; CCVs are indicated by orange arrows.

More than 20% of the third-party bins colocalised with CTCF binding peak(s) and approximately 10% colocalised with region(s) of H3K27ac histone modification (Table 4.4). Compared to the non-third-party target bins (i.e., non-baited, non-CCV- and non-gene-containing bins that were not involved in third-party interactions), this represented a highly significant enrichment for CTCF (5-fold in rCHi-C data, $p < 5 \times 10^{-13}$ and 2-fold in pCHi-C data, $p < 0.002$). For H3K27ac, the enrichment was less consistent with 3- to 5.5-fold in the rCHi-C data ($p < 2 \times 10^{-6}$), 2-fold in the T-47D pCHi-C data ($p = 0.03$) and no significant enrichment in the GS2 pCHi-C data ($p = 0.49$).

Cell line	Unique third-party bins	IPs involving third-party bins (pCHi-C)	IPs involving third-party bins (rCHI-C)	Number of unique combinations	Unique CCV-containing bins	Unique CCVs	CCVs per bin	Unique gene-containing bins	Unique genes
T-47D	174	219	520	652	155	253	1 to 8	69	74
GS2	488	609	1,539	2,135	276	497	1 to 10	135	144

Table 4.3: Summary of third-party interaction peaks called in T-47D and GS2 libraries. Third-party interaction peaks – interaction peaks where a baited smart bin in the pCHi-C data forms an IP with a non-baited target bin that is neither a smart bin nor harbours a CCV (third-party bin) and this same third-party bin forms an IP with a baited bin that harbours at least one CCV in the rCHi-C data. Third-party IPs were identified using the 2kb-binned rCHi-C and pCHi-C libraries generated in T-47D and GS2 cell lines using the Dovetail Genomics Omni-C protocol. Number of unique combinations – number of unique 'CCV bin – third-party bin – gene bin' combinations. Unique CCV-containing bins – number of unique CCV-containing bait fragments in the rCHi-C data that were involved in third-party IPs. Unique gene-containing bins – number of unique baited smart bins in the pCHi-C data that were involved in third-party IPs.

Cell line	CHi-C type	CTCF/H3K27ac	Third-party bins	Third-party bins with a feature		Non-third-party target bins	Non-third-party bins with a feature		Odds ratio	*p* value
T-47D	rCHi-C	CTCF	174	46	26.44%	1,198	81	6.76%	4.95	4.12×10^{-13}
T-47D	rCHi-C	H3K27ac	174	17	9.77%	1,198	23	1.92%	5.52	1.71×10^{-6}
T-47D	pCHi-C	CTCF	174	46	26.44%	17,565	2,971	16.91%	1.77	0.002
T-47D	pCHi-C	H3K27ac	174	17	9.77%	17,565	972	5.53%	1.85	0.03
GS2	rCHi-C	CTCF	488	98	20.08%	1,795	86	4.79%	4.99	1.92×10^{-23}
GS2	rCHi-C	H3K27ac	488	51	10.45%	1,795	64	3.57%	3.15	1.46×10^{-8}
GS2	pCHi-C	CTCF	488	98	20.08%	38,581	4,638	12.02%	1.84	4.19×10^{-7}
GS2	pCHi-C	H3K27ac	488	51	10.45%	38,581	3,688	9.56%	1.10	0.49

Table 4.4: H3K27ac and CTCF enrichment analysis of third-party bins. *p* values were calculated using a Fisher's exact test. H3K27ac – CUT&Tag data generated in T-47D and GS2 cells by other members of the lab. CTCF – dataset compiled using 9 ChIP-seq samples from the ENCODE generated in breast cell types (3 – breast epithelium, 1 – mammary epithelial cells, 1 – mammary fibroblasts, 4 – MCF-7); only consensus peaks (i.e., peaks that were present in at least 8 out of 9 samples) were considered for the analysis. Non-third-party target bins – non-baited, non-CCV- and non-gene-containing bins that did not form third-party IPs.

Interestingly, in 179 out of 652 T-47D third-party interaction peaks (27.5%) the resulting distance between a CCV-containing bin and a gene-containing bin was less than 10 kb (Table 4.5); the same was true for 384 GS2 third-party interaction peaks (18%). In T-47D, these 'short-range' interaction peaks involved 22 genes and 71 CCVs, more than half of which (15 genes and 38 CCVs) were identified based on third-party interaction peaks exclusively (i.e., they did not appear among rCHi-C or pCHi-C 'direct' genes and CCVs). In GS2, 'short-range' third-party interaction peaks involved 34 genes and 108 CCVs, out of which 25 genes and 53 CCVs were only involved in third-party interaction peaks.

Cell line	Total IPs	*cis* < 10kb	*cis* ≥ 10kb	*trans*	Replicated
T-47D	652	179	439	18	16
GS2	2,135	384	1,624	0	127

Table 4.5: Distribution of distances between a CCV-containing bin and a gene-containing bin in third-party interaction peaks. Replicated – interaction peaks where the same CCV- and gene-containing baited bin formed interaction peak(s) with the same third-party target bin(s) in both rCHi-C and pCHi-C data (Figure 4.5).

Comparing 'third-party' genes/CCVs to 'direct' genes/CCVs, 33 and 21 'direct' genes identified in the T-47D rCHi-C and pCHi-C datasets, respectively, were also involved in third-party interaction peaks (Figure 4.1; GS2: 57 and 51 'direct' genes). In GS2, these included *CCND1* and *MYEOV* that formed direct interaction peaks in the GS2 pCHi-C dataset. *CCND1* formed third-party interaction peaks with the same two CCV-containing bins (as in the pCHi-C data) via chr11:69,296,001-69,298,000 and chr11:69,300,001-69,302,000 third-party bins (Figure 4.2). The region to which these third-party bins map colocalised with an H3K27ac peak in our CUT&Tag data in GS2 but not T-47D cells. *MYEOV* formed third-party interaction peaks with the same bin as *CCND1* (encompassing rs476679, rs602690 and rs510754) via two consecutive third-party bins that span the chr11:69,636,001-69,639,966 region (Figure 4.2). This region also colocalised with an H3K27ac peak in the CUT&Tag data. Interestingly, the third-party bins that were involved in mediating third-party interaction peaks involving *CCND1* were located very close to the *MYEOV* promoter (chr11:69,293,018-69,294,877), while the third-party bins that were involved in mediating third-party interaction peaks with *MYEOV* were located very close to the *CCND1* promoter (chr11:69,639,967-69,641,826). Thus, these third-party interaction peaks effectively bring *CCND1*, *MYEOV* and a subset of CCVs together, raising the possibility that *CCND1* and *MYEOV* may be co-regulated

in some way. A bin containing rs476679, rs602690 and rs510754 also formed third-party interaction peaks with *AP000439.2* (via chr11:69,636,001-69,639,966 and chr11:69,296,001-69,298,000 regions) and with *RP11-554A11.7* (via chr11:69,636,001-69,639,966 region).

The remaining genes and CCVs were exclusively associated with third-party interaction peaks in T-47D (39 genes and 117 CCVs) or GS2 (81 genes and 207 CCVs) (Figure 4.1). Most of these genes mapped to the fine-mapping regions where the rCHi-C and/or pCHi-C data identified one or more putative target genes through direct interaction peaks. However, 12 T-47D 'third-party' genes mapped to 7 fine-mapping regions where neither the rCHi-C, nor the pCHi-C data picked up any putative target genes (GS2: 19 genes at 10 regions). These genes are listed in Table 4.6, except three T-47D 'third-party' genes (*ASPG, RAB40B, CAMK2N2*) that were identified based on *trans* third-party interaction peaks. Such genes should be treated with caution, especially in the highly rearranged cells like T-47D, where they are more likely to represent cell type specific artefacts than valid interactions.

Some of these genes and CCVs were involved in interaction peaks, where the same CCV- and gene-containing baited bin formed interaction peak(s) with the same third-party target bin(s) in both rCHi-C and pCHi-C data ('replicated' interaction peaks; Figure 4.5). Two examples are *FGFR2* and *FAM179A* (T-47D) that each exclusively participated in such 'replicated' interaction peaks. These 'replicated' interaction peaks should be interpreted with caution, especially where these are the only kind of interaction peaks that a given gene or CCV formed. Overall, 9 out of 74 T-47D 'third-party' genes (GS2: 25 out of 144 genes) were involved in the 'replicated' interaction peaks, of which for 3 T-47D (*FAM179A*, *FGFR2*, *L3MBTL3*) and 4 GS2 (*ATXN7*, *THOC7*, *CDKAL1*, *ZBTB38*) genes these were the only type of interaction peaks they formed.

Fine-mapping region	Cell line	Gene	CCV(s)
chr2:28447809-29447810	T-47D	FAM179A*	rs12472404*
chr2:171019711-172608243	T-47D	DLX2, DLX2-AS1	rs2016394, rs17726078; rs544674726
chr8:123097926-124097925	T-47D	FAM91A1	rs13281094, rs7014939; rs4595110, rs28651583; rs17349815; rs34838484
chr10:120834389-122089809	T-47D	FGFR2*	rs7899765*
chr14:90874725-91902279	T-47D	C14orf159, RPS6KA5; CCDC88C*	rs2277509; rs11341843*
chr22:39980230-41131866	T-47D	MKL1*	rs12158872*; rs56283550; rs17001907; rs551057361, rs56215843
chr1:204049714-205049714	GS2	NFASC	rs4951401; rs930947, rs4951404
chr2:24464730-25464730	GS2	ADCY3; CENPO, PTRHD1	rs6746013; rs2384057, rs10865315, rs2033655
chr2:119987546-120988992	GS2	INHBB; AC012363.13*; GLI2; AC018866.1	rs13018516*; rs34160433; rs7593535; rs11903787; rs17625845
chr5:32067626-33067626	GS2	SUB1	rs12519859
chr5:169664483-170664483	GS2	FOXI1*	rs56225360*; rs4315934
chr8:104846392-105849165	GS2	ZFPM2	rs2957440, rs56128159
chr8:123097926-124097925	GS2	FBXO32; FAM91A1; RP11-245A18.1	rs17253058; rs547278904; rs35542655; rs34838484; rs4401839, rs4509301
chr14:90874725-91902279	GS2	CCDC88C*; C14orf159, RPS6KA5	rs2277509; rs11341843*
chr17:79794855-79816335	GS2	CBX8*	rs4889891; rs8082452*, rs9905914*
chr19:29286822-30286822	GS2	ZNF536	rs12461902, rs62107106

Table 4.6: 'Third-party' genes that mapped to fine-mapping regions where no genes that formed direct interaction peaks in the rCHi-C or pCHi-C data were identified. 9 T-47D 'third-party' genes mapped to 7 fine-mapping regions where neither the rCHi-C, nor the pCHi-C data picked up any putative target genes (GS2: 19 genes at 10 regions). In the T-47D data there were 3 additional 'third-party' genes which were identified based on *trans* third-party interaction peaks; these were excluded. Fine-mapping region coordinates are in GRCh38/hg38. (*) – genes or CCVs that participated in 'replicated' interaction peaks; in red – genes or CCVs that formed 'replicated' interaction peaks only. CCVs separated with comma map to the same CCV-containing bins, with semicolon – to different bins.

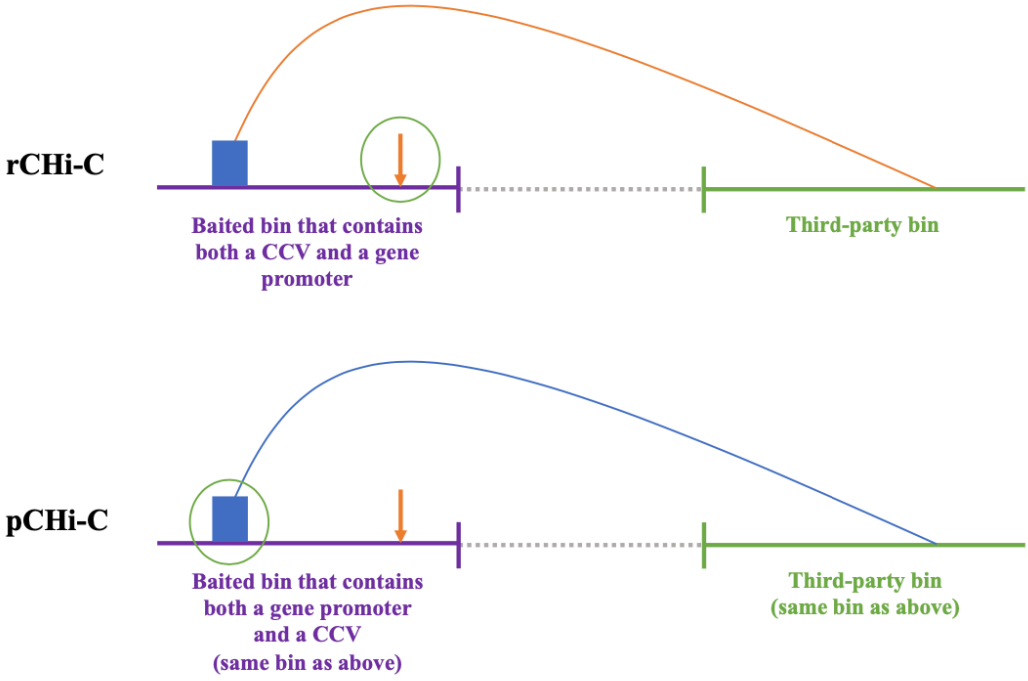

Figure 4.5: Replicated interaction peaks. 'Replicated' interaction peaks are interaction peaks where the same CCV- and gene-containing baited bin formed interaction peak(s) with the same third-party target bin(s) in both rCHi-C and pCHi-C data.

However, some of the 'third-party' genes may represent valid putative targets. For example, there are two independent 'strong-evidence' signals – signal 1 (23 CCVs) and signal 2 (6 CCVs) at the chr8:123,097,926-124,097,925 region (8q24.13 locus). In GS2 cells, the *FBXO32* promoter formed third-party interaction peaks with three signal 2 CCVs – rs35542655, rs17253058 and rs547278904 via two consecutive third-party bins spanning the chr8:123,690,001-123,694,000 region that was found to colocalise with two consensus CTCF sites (Figure 4.6). Two other 'third-party' genes (*FAM91A1* and *RP11-245A18.1*) also formed third-party interaction peaks at this fine-mapping region; however, these interaction peaks only involved signal 1 CCVs. Interestingly, *FBXO32* is an example of a gene that participated in 'short-range' third-party interaction peaks. Although third-party interaction peaks that involved the *FBXO32* promoter are mediated via a region located ~ 150 kb away, the linear distance between the *FBXO32* promoter and the three interacting CCVs is between 6 – 12 kb.

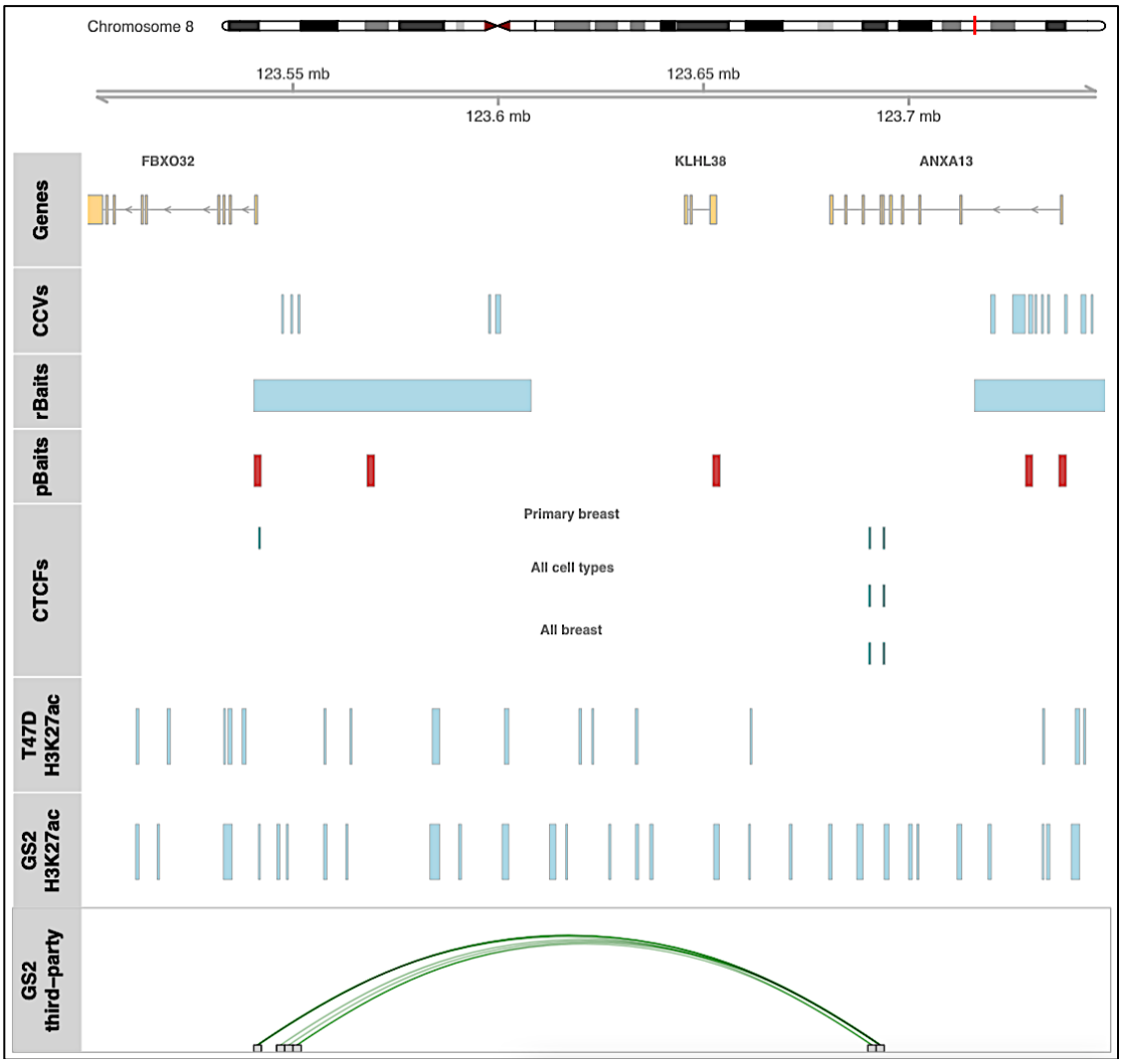

Figure 4.6: Third-party interaction peaks at 8q24.13 involving the *FBXO32* promoter. Third-party interaction peaks identified in GS2 cells at the 8q24.13 breast cancer risk locus (chr8:123,097,926-124,097,925 fine-mapping region, hg38) that involved the *FBXO32* promoter. Two other 'third-party' genes (*FAM91A1* and *RP11-245A18.1*) that also formed third-party interaction peaks at this locus are not shown. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. pBaits – pCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). H3K27ac – H3K27ac peaks identified from CUT&Tag data generated in T-47D and GS2 cell lines.

4.4. Discussion

Generating pCHi-C libraries in addition to the rCHi-C libraries allowed me to examine the extent to which these two approaches prioritise the same set of putative target genes and CCVs, through any direct interaction peaks or identical interaction peaks. Generating both data types also allowed me to investigate the possibility that a target gene and a CCV might be brought into proximity with each other by forming interaction peaks with a 'third party' (rather than by interacting with each other directly).

Overall, many more interaction peaks were called in the pCHi-C compared to rCHi-C datasets. There was, however, a higher proportion of direct interaction peaks (and a higher absolute number of direct interaction peaks) in the rCHi-C data (compared to the pCHi-C data). This difference is probably due to differences in array design. In rCHi-C, baits are designed to capture genetic variants at a limited number of GWAS signals defined by linkage disequilibrium blocks. In pCHi-C, baits are designed to capture annotated gene promoters genome-wide, resulting in a much larger array but with the majority of baits mapping megabases away from the nearest GWAS signal. Specifically, while pCHi-C generates more data, much of that data is not relevant for the purposes of annotating GWAS signals.

The observation that there were more direct interaction peaks as well as 'direct' gene- and CCV-containing bins in the 5kb-binned pCHi-C datasets (compared to the 2kb-binned) suggests that the pCHi-C data lacked power for a 2 kb analysis. Sequencing the pCHi-C data to a greater depth might resolve this issue. The fact that the reverse was true for the rCHi-C data (there were more direct interaction peaks and 'direct' gene- and CCV-containing bins in the 2kb-binned data compared to the 5kb-binned) may reflect the smart bin design of the Dovetail Pan Promoter Enrichment Panel and the Dovetail analysis pipeline. This is explained in more detail in Section 6.1.

Comparing genes and CCVs that were involved in direct interaction peaks identified in each of the two methods, the majority of 'direct' genes (T-47D: 72%; GS2: 80%) and CCVs (T-47D: 62%; GS2: 58%) identified by pCHi-C datasets were also involved in direct interaction peaks in rCHi-C datasets. Comparing 'direct' interaction peaks (rather than 'direct' genes and CCVs individually), almost 60% of the direct interaction peaks identified in pCHi-C data were replicated in the rCHi-C data (Figure 4.3). Almost all 'direct' genes and CCVs that were picked up by both methods were involved in these reciprocal interaction peaks, but some of these genes and CCVs were also involved in method-specific direct interaction peaks.

Some of the genes and CCVs that were 'uniquely' picked up by the pCHi-C mapped to regions that were excluded from my rCHi-C capture array, meaning that I could not pick these genes up using my rCHi-C data. However, a small subset of genes that were 'unique' to the pCHi-C datasets mapped to the regions that were targeted by the rCHi-C array but where there were no 'direct' genes in the rCHi-C data. This difference may be

due to a lack of coverage at the CCVs in the rCHi-C array due to difficulties in designing baits to the relevant region. The overall coverage of my rCHi-C array was 78%, but the percentage varied largely from region to region. As a result, some of the functional CCVs could have mapped to the portions of the regions where the Agilent eArray software was unable to design baits.

Investigating the possibility that a target gene and a CCV might be brought into proximity with each other by forming interaction peaks with a 'third party', I identified 174 (T-47D) and 488 (GS2) third-party bins that were involved in 652 and 2,135 third-party interaction peaks, respectively. These third-party bins were highly enriched for CTCF binding and less enriched for the active histone modification H3K27ac (Table 4.4). The reasons for the differences in the proportions of non-third-party target bins which colocalised with one of these markers in the rCHi-C data compared to the pCHi-C data are not clear, but it may, in part, be explained by how third-party (and non-third-party) target bins were defined and the opposite capture viewpoints used in the two methods. In this thesis, third-party (and non-third-party) target bins were non-baited, non-gene- and non-CCV-containing bins. For H3K27ac, this may have led to differential depletion of peaks in the pCHi-C and rCHi-C data. Specifically, in the pCHi-C data, where, by definition, all interactions originate at a gene promoter, a significant proportion of target bins would be expected to represent active enhancer elements of which only a minority will be tagged by a CCV (and, therefore, excluded). By contrast, in the rCHi-C data, the interactions originate at a CCV-containing bin, and a significant proportion of target bins would be expected to represent active promoters, all of which have been excluded. Since H3K27ac histone modifications are associated with both active enhancers and active promoters, regions of H3K27ac modification may have been more systematically excluded from the rCHi-C data compared to the pCHi-C data.

Interestingly, in a subset (T-47D: 27.5%; GS2: 18%) of third-party interaction peaks the resulting distance between a CCV-containing bin and a gene-containing bin was less than 10 kb (Table 4.5). A relatively large proportion of genes and CCVs that were involved in these 'short-range' third-party interaction peaks were identified based on third-party interaction peaks exclusively (i.e., they did not appear among rCHi-C or pCHi-C 'direct' genes and CCVs). As mentioned previously, the ability of CHi-C to detect interaction peaks over distances of less than 10 kb is limited[115]. Therefore, one possible option for overcoming this limitation may be the use of third-party (indirect) interaction peaks.

Almost half of the genes (and over half of CCVs) that were involved in third-party interaction peaks also formed direct interaction peaks in either rCHi-C or pCHi-C or both, while the other half was exclusively associated with third-party interaction peaks. Since 'indirect' interaction peaks are a relatively new concept, it is unclear whether genes and CCVs that form 'indirect' interaction peaks are any different from those that form direct interaction peaks, and which ones should be prioritised over the others. The examples that I picked as illustrations of third-party interaction peaks (*FBXO32*, *CCND1* and *MYEOV*) are all genes that are likely to be involved in complex regulatory networks to coordinate their expression with cofactors (*FBXO32*) and cell cycle regulation (*CCND1* and *MYEOV*).

FBXO32 – is an F-box protein. F-box proteins are substrate-recognition subunits of Skp1-Cullin1-F-box protein (SCF) E3 ligase complexes. Studies suggest emerging roles of F-box proteins in carcinogenesis, tumour progression, and drug resistance through degradation of their downstream substrates[122]. *FBXO32* was proposed to have tumour-suppressive function in breast cancer by targeting *KLF4*, a zinc-finger transcription factor involved in a large variety of cellular processes, to proteasomal degradation and, therefore, inhibiting breast cancer development[123]. *FBXO32* deficiency in breast cancer cells leads to *KLF4* accumulation and facilitates tumorigenesis both *in vitro* and *in vivo*. Interestingly, *KLF4* has also been identified as 'high confidence' target gene at the chr9:107,041,527-108,633,073 region by the BCAC INQUISIT algorithm[67] and in an in-depth functional annotation study[124]. *KLF4* was also identified as a putative target gene in 2kb rCHi-C datasets generated in GS2, primary fibroblasts and primary luminal epithelial cells.

CCND1 and *MYEOV* map to a known region of cancer-associated amplification at 11q13.3[125]. *CCND1* is an established oncogene that was found to be overexpressed in more than 50% of human breast cancers[126]. Together with its binding partner *CDK4*, *CCND1* acts as a regulator of transcription in the nucleus. A recent study, however, demonstrated that the localisation of CCND1-CDK4 complex in the membrane of normal fibroblasts and tumour cells has an active role in the induction of cell migration and invasion through the phosphorylation of a tyrosine-kinase substrate protein called paxillin, providing an explanation to the invasive properties of *CCND1*-overexpressing tumours[127]. *MYEOV* was found to be amplified in 9.5% of breast tumours (most frequently together with *CCND1*) and abnormally expressed in 16.6% of tumours[128].

Although dysregulated expression of *MYEOV* has been associated with its tumorigenic properties, the molecular mechanisms underlying MYEOV-mediated tumorigenesis are still largely unknown. A recent study suggested that *MYEOV* transcript acts as a competing endogenous RNA (ceRNA) to regulate TGF-β signalling and promote the invasion and metastasis of non-small cell lung cancer cells[129].

Although additional studies are required to confirm the findings, my preliminary results suggest that it is possible that some putative target genes and CCVs might be brought into proximity with each by forming third-party interaction peaks. Third-party bins were enriched for CTCF and, less so, for H3K27ac, suggesting they may have a structural or regulatory role. Investigating third-party interaction peaks may help to overcome one limitation of CHi-C methodology, namely the lack of resolution of this technique for distances of less than 10 kb. It may also pinpoint additional (plausible) putative target genes.

Thus, pCHi-C may represent a useful complementary approach that allows reciprocal replication of rCHi-C findings as well as the investigation of third-party interaction peaks. In this project I limited my pCHi-C analysis to cell lines, however, it would be interesting to extend it to primary cells.

5. Region Capture Hi-C in primary cells

5.1. Overview of the libraries

Libraries were generated in primary breast luminal epithelial cells and fibroblasts isolated from each of two women undergoing reduction mammoplasty (four libraries in total) using the Dovetail Genomics Omni-C protocol. Sequencing data obtained for the primary cell libraries demonstrated the same pattern as that of the cell line libraries, namely the percentage of on-target pairs was higher in libraries that had lower absolute numbers of unique read pairs (Table 5.1). Sequencing data for the cell type replicates were combined for the analysis resulting in two datasets – luminal epithelial (EPI, hereafter) and fibroblast (FIB). Significant interaction peaks were called in the 2kb- and 5kb-binned data using CHiCANE (Table 5.2). The number of interaction peaks varied from 14,036 to 24,421, with larger numbers of interaction peaks called in the EPI dataset compared to the FIB dataset. Higher proportions of *cis* ≥ 1 Mb interactions (EPI: 60% vs. 14%; FIB: 75% vs. 19%) were again observed in the 5kb-binned datasets, supporting the idea that size of the baited bins influences the distance range within which the majority of interaction peaks are called.

5.2. Overview of all interaction peaks

As with the cell line data, the primary cell rCHi-C data were mapped to: (i) 84,643 promoters associated with 27,375 coding and non-coding genes; (ii) 5,117 CCVs associated with 196 ‘strong-evidence’ breast cancer risk GWAS signals.

The proportion of unique gene-containing bins out of the total unique bins in each given dataset varied from 2.0% to 4.9% (Table 5.3). There were 112 to 481 unique genes that participated in 287 to 3,078 interaction peaks with a median of 2 to 3 interaction peaks per individual gene. The proportion of unique CCV-containing bins was between 6.1% and 6.5%. These CCV-containing bins harboured a total of 957 to 1,260 unique CCVs (18.7% to 24.6% out of 5,117 CCVs) that participated in 5,630 to 6,830 interaction peaks.

Dataset	Sample	Total pairs	Unique pairs		On-target pairs		*cis* pairs		*cis* ≤ 1kb		*cis* 1kb - 10kb		*cis* 10kb - 1Mb		*cis* > 1Mb	
EPI	3002N	785,271,324	367,045,146	47%	40,310,878	11%	35,064,599	87%	9,166,591	26%	6,175,474	18%	13,784,780	39%	5,937,754	17%
EPI	1989N	655,771,355	305,499,839	47%	39,638,228	13%	33,528,502	85%	5,526,226	16%	5,891,266	18%	15,481,965	46%	6,629,045	20%
FIB	3002N	821,260,358	442,550,078	54%	35,840,961	8%	26,801,390	75%	4,945,724	18%	2,943,321	11%	10,743,195	40%	8,169,150	30%
FIB	1989N	380,473,099	216,611,521	57%	43,760,897	20%	34,213,837	78%	5,999,098	18%	5,133,706	15%	14,598,703	43%	8,482,330	25%

Table 5.1: Summary sequencing statistics for primary luminal epithelial and fibroblast rCHi-C libraries. Summary sequencing statistics for the rCHi-C libraries generated in primary breast luminal epithelial cells and fibroblasts isolated from each of two women (samples 3002N and 1989N) undergoing reduction mammoplasty using the Dovetail Genomics Omni-C protocol. Total pairs – total number of read pairs where both ends aligned uniquely to the reference genome. On-target pairs – read pairs for which at least one end overlaps with a capture array probe (minimum overlap = 1 bp).

Dataset	Bin size	Total IPs	*trans* IPs		*cis* < 1kb		*cis* 1kb - 10kb		*cis* 10kb - 100kb		*cis* 100kb - 1Mb		*cis* ≥ 1Mb	
EPI	5kb	14,833	281	1.9%	0	0.00%	8	0.05%	162	1.1%	5,520	37%	8,862	60%
EPI	2kb	24,421	103	0.4%	11	0.05%	27	0.11%	1,004	4.1%	19,847	81%	3,429	14%
FIB	5kb	14,036	86	0.6%	0	0.00%	1	0.01%	76	0.5%	3,329	24%	10,544	75%
FIB	2kb	18,373	38	0.2%	9	0.05%	16	0.09%	550	3.0%	14,260	78%	3,500	19%

Table 5.2: Interaction peak calling statistics for EPI and FIB rCHi-C datasets. A breakdown of significant (q-value ≤ 0.1) interaction peaks called in the rCHi-C datasets generated in primary luminal epithelial cells and fibroblasts using the Dovetail Genomics Omni-C protocol and called using CHiCANE is shown. Each dataset was called using 2kb- and 5kb-binned data.

5.3. Direct interaction peaks

Next, I focused on interaction peaks in which a bin colocalising with a gene promoter formed a direct interaction with a CCV-containing bin (Table 5.4). The number of direct interaction peaks varied from 66 to 381. They involved 36 to 158 unique gene-containing bins harbouring a total of 39 to 157 'direct' genes (1 to 3 genes per bin). Overall, 117 to 429 CCVs were involved in direct interaction peaks ('direct' CCVs). On average, larger numbers of CCVs per bin were observed in the 5kb datasets rather than in 2 kb datasets (EPI: 1 to 13 vs. 1 to 10; FIB: 1 to 10 vs. 1 to 8). There were larger numbers of individual genes and CCVs participating in direct interaction peaks in the 2kb datasets rather than 5kb datasets.

Comparing similarity across bin sizes, 50 out of 64 genes found in the 5kb EPI dataset appeared in the 2kb dataset (Figure 5.1; FIB: 32 out of 39 genes). As expected, most of the 'direct' genes identified using the 5kb- but not 2kb-binned datasets formed longer-range interaction peaks (EPI: 1.5 Mb – 5 Mb; FIB: 1.7 Mb – 5.8 Mb). The only exception was *CDCA7* gene that formed one direct interaction peak with the interaction distance of ~ 22 kb in both 5kb EPI and FIB datasets.

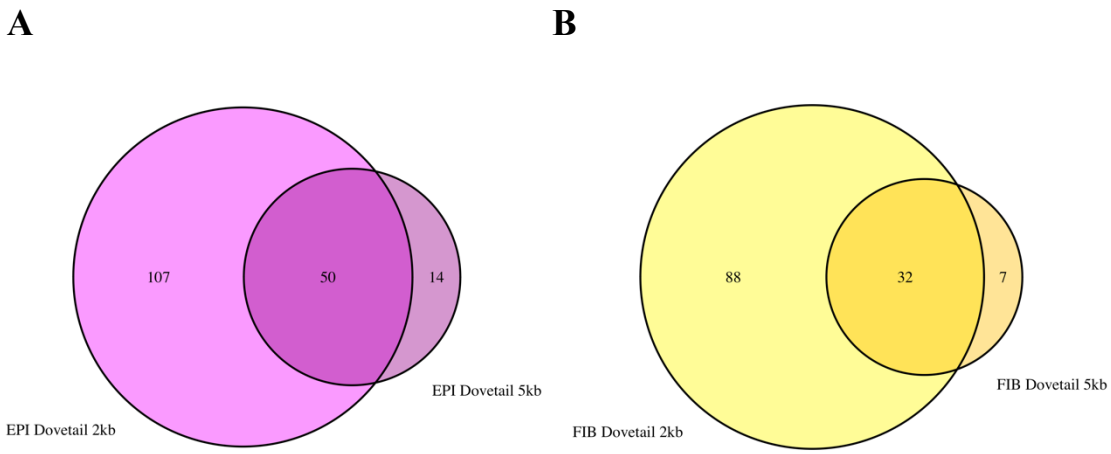

Figure 5.1: Venn diagrams illustrating the overlap between 'direct' genes identified in the 2kb- and 5kb-binned Dovetail rCHi-C libraries generated in primary luminal epithelial cells and fibroblasts. (A) EPI Dovetail 5kb and EPI Dovetail 2kb datasets; (B) FIB Dovetail 5kb and FIB Dovetail 2kb datasets.

Dataset	Bin size	Total IPs	Total unique bins in a dataset	Total unique gene-containing bins		Total unique genes	Number of IPs involving gene-containing bins	IPs per unique gene (median + range)	Total unique CCV-containing bins		Total unique CCVs (out of 5,117)		Number of IPs involving CCV-containing bins
EPI	5kb	14,833	6,863	173	2.5%	179	595	2 (1 to 34)	444	6.5%	1,234	24.1%	6,207
EPI	2kb	24,421	10,231	498	4.9%	481	3,078	3 (1 to 148)	632	6.2%	1,260	24.6%	6,830
FIB	5kb	14,036	5,743	112	2.0%	112	287	2 (1 to 14)	366	6.4%	1,028	20.1%	5,873
FIB	2kb	18,373	8,082	338	4.2%	318	1,798	3 (1 to 65)	496	6.1%	957	18.7%	5,630

Table 5.3: Summary of interaction peaks identified in primary luminal epithelial and fibroblast rCHi-C datasets for which the interacting fragments colocalised with: (i) an annotated RefSeq gene promoter; (ii) one or more CCVs selected by the BCAC fine-scale mapping analysis.

Dataset	Bin size	Total direct IPs	Unique gene-containing bins	Unique genes	Unique CCV-containing bins	Unique CCVs	CCVs per bin
EPI	5kb	97	57	64	63	198	1 to 13
EPI	2kb	381	158	157	224	429	1 to 10
FIB	5kb	66	36	39	38	117	1 to 10
FIB	2kb	278	122	120	157	317	1 to 8

Table 5.4: Summary of direct interaction peaks called in primary luminal epithelial and fibroblast rCHi-C datasets. Direct interaction peaks – interaction peaks in which a bin colocalising with a gene promoter forms a direct interaction with a CCV-containing bin.

In the absence of a ground truth, i.e., a dataset in which the true target genes and causal variants are known, it is difficult to know which is the 'best' analysis. However, as previously, on the grounds that very long-range (> 2 Mb) 'functional' interactions are less plausible than shorter-range ones, and with a view to maximise resolution, the rest of the analysis is focused on the 2kb-binned datasets only.

5.4. Prioritisation of putative target genes

To evaluate the ability of rCHi-C to prioritise putative target genes at breast cancer risk loci, 'direct' genes were mapped back to 129 'strong-evidence' breast cancer risk-associated regions (Table 5.5). This resulted in at least one putative target gene identified at 57 regions (1 to 13 genes per region; median = 2) using EPI data, and at 43 regions (1 to 9 genes per region; median = 2) using FIB data. The majority of these regions (42 EPI (73.7%) and 31 FIB (72.1%) regions) contained 1 to 3 putative target genes, which realistically is the number of genes for which in-depth functional follow up studies could be performed.

Next, I compared my sets of putative target genes against 191 genes that were predicted with 'high confidence' using the BCAC's integrated-expression quantitative trait and in silico prediction of GWAS targets algorithm (INQUISIT)[67]. Overall, INQUSIT predicted at least one 'high confidence' target gene at 88 fine-mapping regions. Since 7 out of 129 regions were not covered by the capture array, they were excluded from further analysis, bringing the number of 'high confidence' INQUISIT genes to 177, and the number of corresponding regions to 84 (Table 5.5).

Out of 177 INQUISIT genes, 18 were predicted at the same fine-mapping regions by rCHi-C in both EPI and FIB datasets, 17 were predicted in just the EPI dataset and 4 were predicted in just the FIB dataset, suggesting a greater concordance between the INQUISIT predictions and those from the EPI data compared to those from the FIB data. There are 38 regions where INQUISIT did not predict any 'high-confidence' genes; at 12 of these regions rCHi-C data in EPI and/or FIB predicted at least one gene (Table 5.5). These include several plausible candidate genes, such as *OLA1* (at chr2:172,846,180-173,848,166) in EPI and *FGF10* (at chr5:44,013,202-45,206,396) in FIB.

Fine-mapping region	EPI genes	FIB genes	INQUISIT genes
chr1:9983762-11006158	APITD1; APITD1-CORT; TARDBP		CASZ1; PEX14
chr1:17980845-18980845			KLHDC7A
chr1:45635245-46635245	NSUN4		LRRC41; MAST2; PIK3R3; POMGNT1
chr1:87191240-88191240	LMO4; LINC01140	LMO4; LINC01140	
chr1:113405767-114405767	HIPK1		RSBN1
chr1:145625092-146010623*	NUDT17		NUDT17; PDZK1; PIAS3; POLR3GL; RNF115
chr1:154676305-155678990	MTX1; THBS3		EFNA1; FAM189B; GBA; MTX1; MUC1; RP11-263K19.4; SLC50A1; THBS3; TRIM46
chr1:200968704-201968704			TNNI1
chr1:203332121-204332121	ETNK2; SOX13	ETNK2; SOX13	SOX13; ZC3H11A
chr1:204049714-205049714			LRRN2; MDM4; PIK3C2B; PPP1R15B
chr1:241360598-242370961			EXO1
chr2:18634525-19621042			OSR1
chr2:24464730-25464730			ADCY3; DNMT3A
chr2:28447809-29447810	ALK; CLIP4	ALK; YPEL5; CLIP4	ALK; PPP1CB; TRMT61B
chr2:119987546-120988992			INHBB
chr2:171019711-172608243	DLX2; DLX2-AS1	DCAF17; METTL8; DLX2; DLX2-AS1; ITGA6	DYNC1I2
chr2:172846180-173848166	OLA1; LINC01305	LINC01305	
chr2:200816524-201816524	CFLAR	CASP10; CFLAR; FAM126B; NDUFB3; NIF3L1	ALS2CR12; CASP8; CFLAR; NIF3L1; PPIL3
chr2:216541109-217931785		IGFBP5	IGFBP5
chr3:26786474-28243756	CMC1; AZI2; ZCWPW2; RP11-222K16.2	CMC1; AZI2; ZCWPW2	
chr3:30134389-31134388			TGFBR2
chr3:46325375-47339998			CCDC12; NBEAL2
chr3:63456021-64482224	ATXN7; THOC7; PRICKLE2; RP11-14D22.1; RP11-14D22.2; ABHD14B; ACY1	PRICKLE2; RP11-14D22.1; RP11-14D22.2	ATXN7; PSMD6; PSMD6-AS2; THOC7
chr3:86488393-87488393	CGGBP1; ZNF654	CGGBP1; ZNF654	
chr3:99504736-100504736			CMSS1; FILIP1L; TBC1D23
chr3:140894017-141894017		ZBTB38	ZBTB38
chr3:172067447-173067447			TNFSF10
chr4:38311255-39311256	TLR1; TLR6		TLR1
chr4:82948971-83948971	NKX6-1	NKX6-1; HNRNPD	MRPS18C
chr4:104647856-105935604	PPA2; GSTCD; INTS12; RP11-556I14.2; CXXC4; UBE2D3	TET2; PPA2; GSTCD; INTS12	AC004066.3; ARHGEF38; TET2
chr5:779675-1797374			CLPTM1L; TERT

chr5:44013202-45206396	NNT	PAIP1; NNT; FGF10; RP11-53O19.3; C5orf34	
chr5:56236057-57292056	MAP3K1; C5orf67; CTD-2227I18.1; CTC-236F12.4; ANKRD55; RP11-155L15.1; IL6ST; AC008914.1; PLK2	MAP3K1; RP11-155L15.1; CTC-236F12.4; AC008914.1; IL6ST	MAP3K1
chr5:58388234-59569743		PDE4D	
chr5:90993653-91993653	RP11-213H15.1; RP11-414H23.3; RP11-414H23.2; MBLAC2; POLR3G; ADGRV1; LYSMD3; MCTP1	RP11-213H15.1; RP11-414H23.3; RP11-414H23.2	ARRDC3
chr5:132571366-133571367	SKP1; C5orf15; CTB-113I20.2; VDAC1; WSPAR	C5orf15	AFF4; HSPA4; ZCCHC10
chr5:158303005-159317075	CTC-436K13.1; CTC-436K13.5	CTC-436K13.1; CTC-436K13.5	EBF1
chr5:169664483-170664483			FOXI1
chr6:13212867-14222292			NOL7; RANBP9; RP1-223E5.4
chr6:20121007-21121007	E2F3; CASC15; SOX4; DCDC2; RP11-524C21.2	CASC15; E2F3; SOX4; DCDC2; RP11-204E9.1; NBAT1	CDKAL1
chr6:80918669-82086234	FAM46A	FAM46A	
chr6:129527974-130527974		EPB41L2; SMLR1; AKAP7	L3MBTL3
chr6:151097720-152615881	MTRF1L; ESR1	ESR1	C6orf211; CCDC170; ESR1
chr7:93984487-94984487	PEG10; SGCE; CASD1; AC004012.1; PPP1R9A; ASB4; PON3; PON2; AC002429.5	PEG10; SGCE; PPP1R9A; PON2; AC004012.1; PON3; AC002429.5; ASB4	
chr7:101411318-102414152			CUX1
chr8:60001-720692	DLGAP2		
chr8:29152099-30152100	DUSP4	DUSP4	DUSP4
chr8:36500965-37501668	ERLIN2		ZNF703
chr8:74818066-76005702	PEX2		
chr8:100966731-101966731	GRHL2	NCALD	GRHL2; KB-1562D12.1; KB-1930G5.4
chr8:115697322-116697309	TRPS1		TRPS1
chr8:123097926-124097925	FBXO32		ATAD2; FBXO32; WDYHV1
chr8:126412414-129029685	CASC11; MYC; PVT1; PCAT1; RP11-419K12.1	CASC11; MYC; RP11-419K12.1; PVT1; LINC00976; CCDC26	MYC
chr9:107041527-108633073	KLF4	KLF4; LINC01509	KLF4
chr10:8546150-9546150	GATA3; GATA3-AS1; RP11-379F12.4; ATP5C1; KIN; TAF3	GATA3; GATA3-AS1; RP11-379F12.4	GATA3
chr10:21244013-22620463	BMI1	MSRB2	BMI1; COMMD3
chr10:78581391-79627965		RPS24	ZCCHC24; ZMIZ1
chr10:112514168-113526395	TCF7L2; VTI1A; ZDHHC6; ACSL5	VTI1A; ZDHHC6; TCF7L2; ACSL5	TCF7L2
chr10:120834389-122089809	RP11-95I16.2; RP11-95I16.6		FGFR2

chr11:800484-822622*			CD151; EPS8L2; HRAS; PDDC1; PIDD1
chr11:1377434-2421345		TNNT3	
chr11:65276356-66315595	AP5B1		AP5B1; CFL1; KAT5; OVOL1; RNASEH2C
chr11:69509114-69521223*			CCND1; MYEOV
chr11:107974789-108986410			ATM; C11orf65; KDELC2
chr11:129082612-130091276	BARX2		BARX2
chr12:13760997-14760997	ATF7IP	ATF7IP	ATF7IP
chr12:27486913-28881482	PTHLH	PTHLH	CCDC91; PTHLH
chr12:95133983-96133981	NTN4		NTN4
chr12:114898717-115898717	TBX3; RP11-162N7.1; RP11-116D17.4; RP4-601P9.2	TBX3; RP11-110L15.1; RP11-162N7.1; RP4-601P9.1; RP11-411G2.2; RBM19	TBX3
chr12:119894342-120894343	MSI1; PXN	PXN	MSI1; RPLP0
chr13:31894673-32898488			BRCA2
chr13:72890381-73890382			KLF5
chr14:36163563-37166547	SLC25A21; MIPOL1; TTC6	SLC25A21; MIPOL1	PAX9; SLC25A21
chr14:67650477-69067965	ZFP36L1	ZFP36L1	ZFP36L1
chr14:92137728-93150006			RIN3
chr15:90465983-91465985			RCCD1
chr16:3556787-4556787			ADCY9; CREBBP
chr16:52004913-53004913			TOX3
chr16:53267042-54321379	CRNDE; IRX5; CTD-3032H12.2; CTD-3032H12.1; IRX3; AC007491.1; RP11-324D17.1; IRX6; RP11-26L20.5; AMFR; FTO; NOD2; MMP2	CRNDE; IRX5; CTD-3032H12.2; CTD-3032H12.1; IRX3; RP11-212I21.2; RP11-212I21.5; MMP2; AC007491.1	CTD-3032H12.2; FTO; IRX3
chr16:54148152-55148152	CRNDE; IRX5; IRX3; CTD-3032H12.1; CTD-3032H12.2; AKTIP; RP11-434E6.4		CRNDE
chr16:80114430-81117200	WWOX		CDYL2; DYNLRB2; RP11-18F14.4; RP11-525K10.3
chr16:86551631-87551631	FOXL1; FOXC2; MTHFSD	FOXC2; FOXL1; MTHFSD	
chr17:30384649-31403502			NF1
chr17:79794855-79816335*			CCDC40
chr18:26252512-27495432			KCTD1
chr18:44319625-45319625	SLC14A1; RP11-456K23.1; SETBP1		SETBP1
chr19:12547463-14343759	IER2; STX10; LYL1; NACC1; TRMT1	LYL1; IER2; STX10; NACC1; TRMT1	CACNA1A; DAND5; GCDH; HOOK2; JUNB; MAST1; PRDX2
chr19:16684212-17783315	HAUS8; MYO9B		ABHD8; ANKLE1; MRPL34

chr19:17939625-18960332	HOMER3		CRLF1; ELL; FKBP8; GDF15; ISYNA1; KXD1; UBA52
chr19:29286822-30286822			CCNE1
chr19:43279295-44282360			KCNN4; PLAUR; SMG9
chr21:14701662-15701664			NRIP1
chr22:27604039-29725488			CHEK2; CTA-292E10.6; EWSR1; XBP1
chr22:37672826-39463350			CBX6; MAFF; NPTXR; PLA2G6; SUN2; TMEM184B
chr22:39980230-41131866	MKL1; ST13; XPNPEP3	MKL1	MKL1; SLC25A17
chr22:41142782-42142785			EP300

Table 5.5: Prioritisation of putative target genes at 129 breast cancer risk regions using primary cell rCHi-C data. Out of 129 fine-mapping regions to which 196 'strong-evidence' breast cancer risk signals map, there were 57 at which 157 unique genes formed direct IPs in luminal epithelial cells (EPI genes), and 43 at which 120 unique genes formed direct IPs in fibroblasts (FIB genes). INQUISIT genes – 191 putative target genes predicted with 'high confidence' by Fachal and colleagues at 88 fine-mapping regions using INQUISIT algorithm (reduced to 177 putative target genes at 84 regions after excluding 7 regions that were not covered by the capture array). Regions where no putative target genes were identified using primary cell rCHi-C data or INQUISIT prediction are not shown. INQUISIT genes that were predicted at the same fine-mapping regions using both EPI and FIB rCHi-C datasets are in red, only EPI dataset – in blue, only FIB dataset – in green. Fine-mapping region coordinates are in GRCh38/hg38. (*) – there were several fine-mapping regions (originally defined in hg19) that when lifted over to hg38 were split or partially deleted in hg38. These regions were compiled manually to encompass all CCVs at each of the regions.

5.5. Prioritisation of risk-associated variants

To assess the level to which rCHi-C data can help to prioritise risk-associated variants at GWAS risk loci, 'direct' CCVs were mapped back to 196 'strong-evidence' breast cancer risk signals. Out of 183 signals at which my capture array covered at least one CCV, there were 79 at which at least one CCV formed direct interaction peaks with at least one gene (Table 5.6). As with the cell line rCHi-C data, there was one additional signal (signal 2 at the chr12:27,486,913-28,881,482 region) which was not targeted by the rCHi-C capture array, but some of its CCVs (that mapped by chance to the non-baited interacting bin) formed direct interaction peaks with the *PTHLH* promoter.

At 57 signals my capture array covered all reported CCVs. For most of these signals, interaction peaks in the rCHi-C data involved only a subset of the CCVs, potentially narrowing down the number that would be prioritised for the follow up studies. For example, out of 13 CCVs at signal 1 of chr2:172,846,180-173,848,166 region (2q31.1 locus; Figure 5.2), my approach prioritised a single CCV (rs930313) that formed two direct interaction peaks in the EPI data (with *OLA1* and *LINC01305*) and one direct interaction peak in the FIB data (with *LINC01305* only). Since this CCV maps ~ 500 bp upstream of *CDCA7*, the CCV-containing bin involved in these interaction peaks also colocalised with the promoter of this gene, potentially implicating *CDCA7* as a putative target gene.

Another example is the region at chr1:203,332,121-204,332,121 (1q32.1 locus; signal 1) that comprises 56 CCVs. Of these, 6 CCVs (rs7520079, rs16852420, rs56395476, rs6664515, rs72745792 and rs67087079) and 7 CCVs (same + rs12026395) formed a total of 9 and 16 direct interaction peaks in FIB and EPI datasets, respectively (Figure 5.3). In epithelial cells, rs12026395 formed a single direct interaction peak with the *ETNK2* promoter, while the remaining 6 CCVs formed direct interaction peaks with both *ETNK2* and *SOX13* promoters. rs7520079 formed two direct interaction peaks with *ETNK2* and three direct interaction peaks with *SOX13*, a bin containing rs56395476 and rs6664515 formed a single interaction peak with *ETNK2* and three interaction peaks with *SOX13*, while rs16852420 and a bin containing rs72745792 and rs67087079 formed two and one direct interaction peaks, respectively, with each of the two genes. In fibroblasts, rs7520079 and the bin containing rs72745792 and rs67087079 each formed a single direct interaction peak with the *ETNK2* promoter only, while the remaining three CCVs formed

direct interaction peaks with both *ETNK2* and *SOX13* promoters (rs16852420 – two interaction peaks with each of the two genes; the bin containing rs56395476 and rs6664515 – two and one interaction peaks with *ETNK2* and *SOX13* respectively).

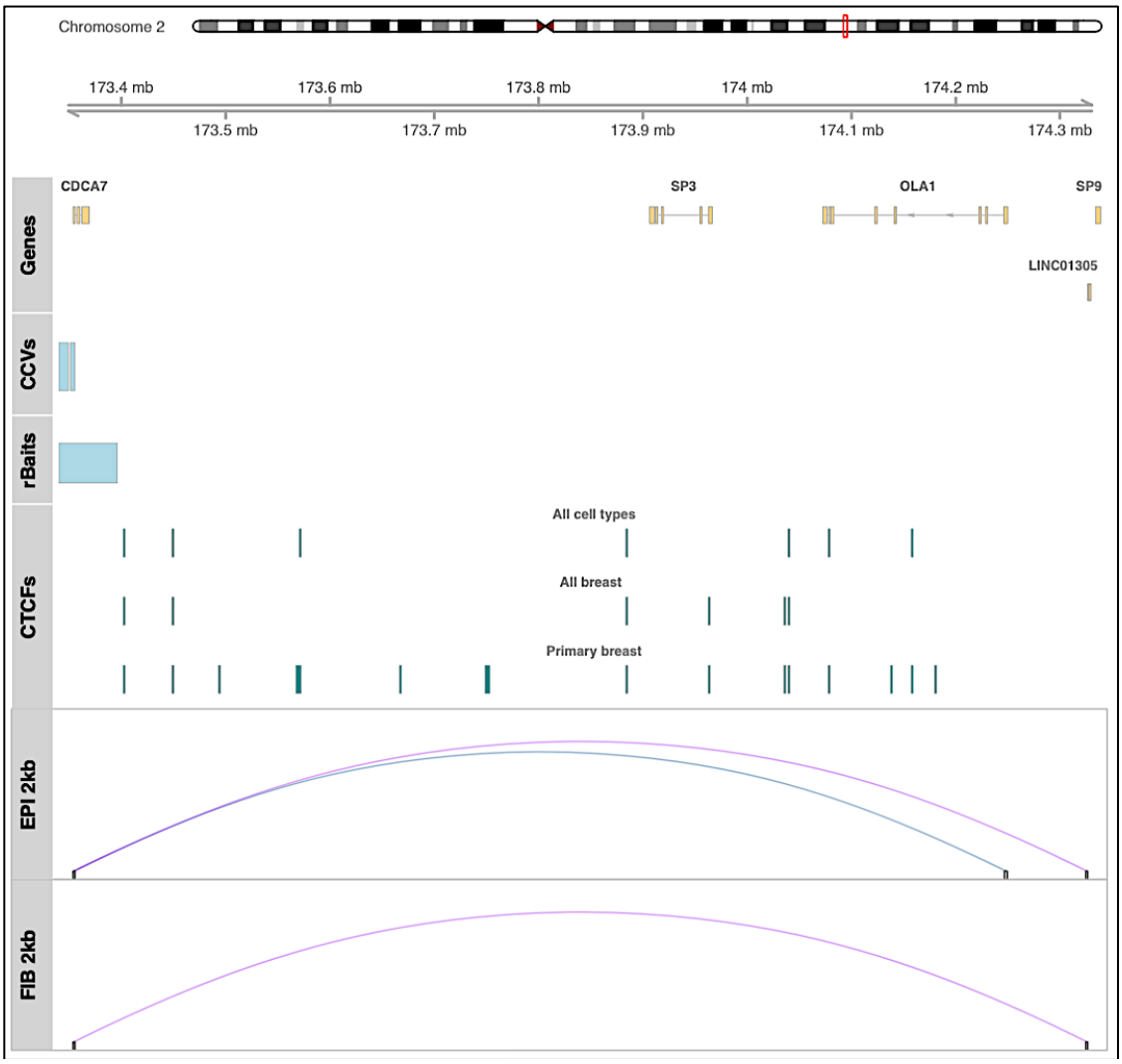

Figure 5.2: Direct interaction peaks at 2q31.1 in EPI and FIB datasets. Direct interaction peaks (shown in a looping format) at the 2q31.1 breast cancer risk locus (chr2:172,846,180-173,848,166 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast luminal epithelial cells (EPI) and fibroblasts (FIB) using the Dovetail Genomics Omni-C protocol. Blue loops – direct IPs that involved the *OLA1* promoter. Purple loops – direct IPs that involved the *LINC01305* promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

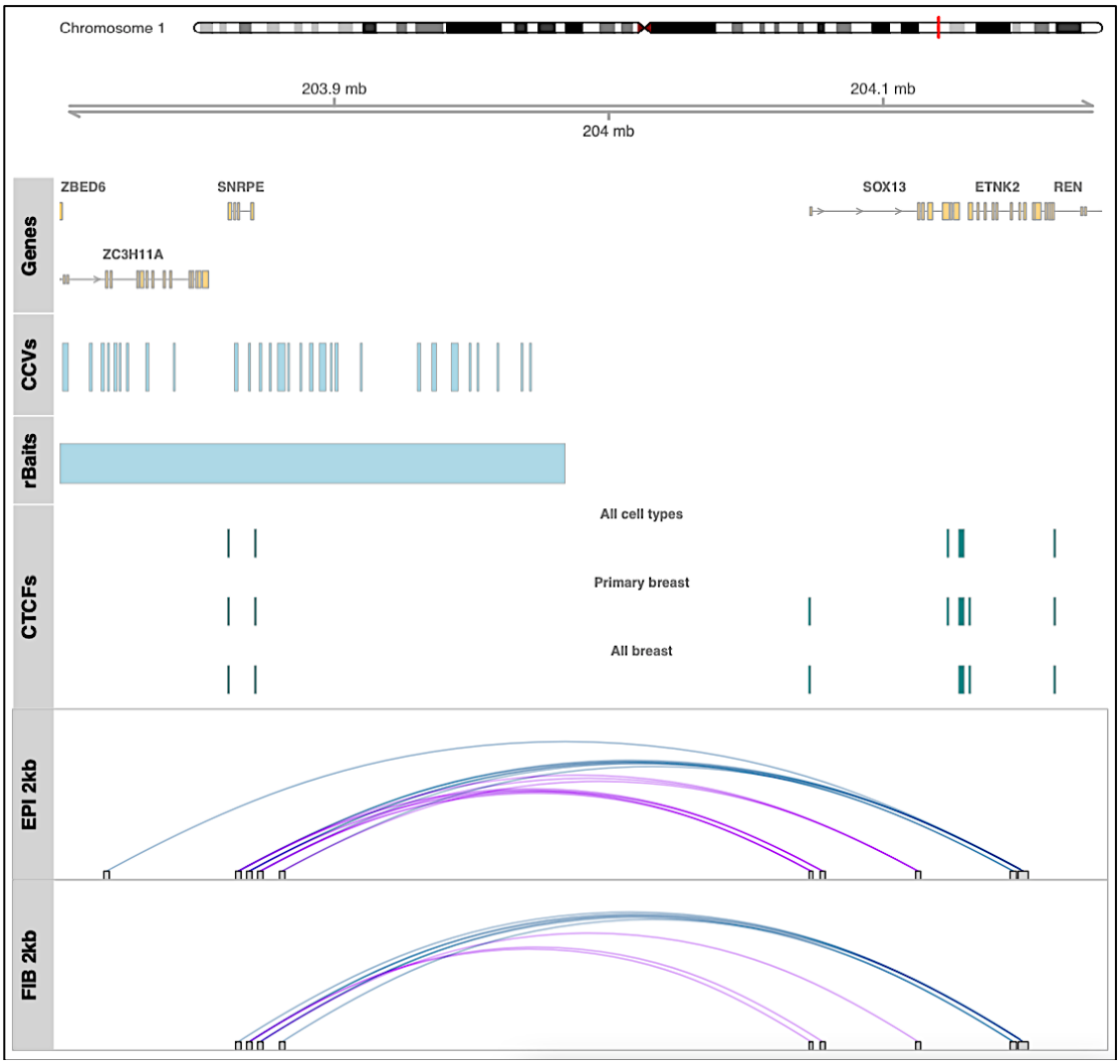

Figure 5.3: Direct interaction peaks at 1q32.1 in EPI and FIB datasets. Direct interaction peaks (shown in a looping format) at the 1q32.1 breast cancer risk locus (chr1:203,332,121-204,332,121 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast luminal epithelial cells (EPI) and fibroblasts (FIB) using the Dovetail Genomics Omni-C protocol. Blue loops – direct IPs that involved the *ETNK2* promoter. Purple loops – direct IPs that involved the *SOX13* promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

Fine-mapping region	Signal	Index SNP	Number of CCVs	Captured CCVs	EPI	FIB
chr1:9983762-11006158	Signal 1	rs657244	19	1	1	0
chr1:45635245-46635245	Signal 1	rs12039667	11	7	1	0
chr1:87191240-88191240	Signal 2	rs11583393	8	8	5	8
chr1:113405767-114405767	Signal 1	rs11102701	12	12	3	0
chr1:145625092-146010623*	Signal 2	rs200366104	5	1	1	0
chr1:154676305-155678990	Signal 1	rs1057941	16	16	1	0
chr1:203332121-204332121	Signal 1	rs59867004	56	56	7	6
chr2:28447809-29447810	Signal 1	rs71403627	81	81	9	1
chr2:171019711-172608243	Signal 2	rs13020413	35	35	1	3
chr2:172846180-173848166	Signal 1	rs7589172	13	13	1	1
chr2:200816524-201816524	Signal 2	rs13015648	8	6	3	2
chr2:216541109-217931785	Signal 2	rs138522813	5	5	0	2
chr2:216541109-217931785	Signal 3	rs5838651	42	42	0	17
chr3:26786474-28243756	Signal 1	rs1352944	44	44	10	4
chr3:26786474-28243756	Signal 2	rs36078735	12	12	2	1
chr3:63456021-64482224	Signal 1	rs555060306	94	94	10	2
chr3:86488393-87488393	Signal 1	rs13066793	2	2	1	1
chr3:140894017-141894017	Signal 1	rs7625643	24	24	0	4
chr4:38311255-39311256	Signal 1	rs10034903	24	24	5	0
chr4:82948971-83948971	Signal 1	rs6854739	84	84	17	16
chr4:104647856-105935604	Signal 1	rs17617028	21	20	4	4
chr5:44013202-45206396	Signal 3	rs13153426	72	65	2	11
chr5:56236057-57292056	Signal 3	rs112497245	21	21	9	4
chr5:56236057-57292056	Signal 4	rs7730210	70	38	14	3
chr5:58388234-59569743	Signal 2	rs10472097	5	1	0	1
chr5:90993653-91993653	Signal 1	rs1964292	88	88	13	8
chr5:132571366-133571367	Signal 1	rs571173399	117	117	24	4
chr5:158303005-159317075	Signal 1	rs31864	5	5	4	2
chr6:20121007-21121007	Signal 1	rs2328531	52	36	3	2
chr6:80918669-82086234	Signal 1	rs7763102	51	41	2	3
chr6:129527974-130527974	Signal 1	rs6569648	43	43	0	16
chr6:151097720-152615881	Signal 2	rs34133739	1	1	1	0
chr6:151097720-152615881	Signal 4	rs7763637	6	6	3	0
chr6:151097720-152615881	Signal 5	rs79388591	173	173	0	20
chr6:151097720-152615881	Signal 6	rs9918437	22	6	1	0

chr7:93984487-94984487	Signal 1	rs1879854	47	42	18	17
chr8:60001-720692	Signal 1	rs34810249	25	24	1	0
chr8:29152099-30152100	Signal 1	rs7465364	16	16	6	2
chr8:36500965-37501668	Signal 1	rs4286946	16	16	1	0
chr8:74818066-76005702	Signal 3	rs17303163	16	16	1	0
chr8:100966731-101966731	Signal 1	rs7813150	47	47	5	6
chr8:115697322-116697309	Signal 1	rs10641009	164	9	3	0
chr8:115697322-116697309	Signal 2	rs13267382	5	2	1	0
chr8:123097926-124097925	Signal 2	rs58847541	6	6	1	0
chr8:126412414-129029685	Signal 2	rs7017073	44	44	23	38
chr8:126412414-129029685	Signal 4	rs419018	43	43	2	0
chr9:107041527-108633073	Signal 1	rs659713	10	10	0	2
chr9:107041527-108633073	Signal 3	rs10816625	1	1	1	0
chr9:107041527-108633073	Signal 4	rs13294895	1	1	1	0
chr10:8546150-9546150	Signal 1	rs7081544	49	48	33	6
chr10:21244013-22620463	Signal 1	rs7098100	7	7	1	0
chr10:21244013-22620463	Signal 2	rs138026227	58	16	0	1
chr10:78581391-79627965	Signal 2	rs10762851	17	17	0	5
chr10:112514168-113526395	Signal 1	rs12250948	12	12	6	4
chr10:112514168-113526395	Signal 2	rs71973726	42	41	1	0
chr10:120834389-122089809	Signal 4	rs2981578	3	3	2	0
chr11:1377434-2421345	Signal 1	rs620315	7	7	0	2
chr11:65276356-66315595	Signal 1	rs548082010	13	12	3	0
chr11:129082612-130091276	Signal 1	rs745382	17	17	1	0
chr12:13760997-14760997	Signal 1	rs12422552	18	18	14	9
chr12:27486913-28881482	Signal 2	rs1600346	375	0	44	16
chr12:95133983-96133981	Signal 1	rs17356907	2	2	2	0
chr12:114898717-115898717	Signal 3	rs1882155	8	8	5	6
chr12:114898717-115898717	Signal 4	rs11067765	6	6	5	5
chr12:119894342-120894343	Signal 1	rs184486140	5	5	3	1
chr14:36163563-37166547	Signal 1	rs12881240	19	19	10	5
chr14:67650477-69067965	Signal 1	rs35378451	8	8	4	6
chr14:67650477-69067965	Signal 2	rs2478777	4	4	2	2
chr16:53267042-54321379	Signal 1	rs55872725	6	6	6	6
chr16:53267042-54321379	Signal 2	rs9925952	21	21	17	11
chr16:54148152-55148152	Signal 1	rs28539243	3	3	3	0
chr16:80114430-81117200	Signal 1	rs9938021	14	14	2	0

chr16:86551631-87551631	Signal 1	rs4066743	85	85	12	18
chr18:27321240-28321240	Signal 1	rs12970390	44	39	0	0
chr18:44319625-45319625	Signal 1	rs78955132	38	38	5	0
chr19:12547463-14343759	Signal 1	rs78269692	21	8	4	1
chr19:16684212-17783315	Signal 1	rs67397200	16	16	1	0
chr19:17939625-18960332	Signal 1	rs8105994	56	56	2	0
chr19:29286822-30286822	Signal 1	rs17513613	60	47	0	0
chr22:39980230-41131866	Signal 1	rs66987842	196	196	18	2

Table 5.6: Prioritisation of CCVs at 196 breast cancer risk signals using primary cell rCHi-C data. 79 signals at which my capture array covered at least one CCV and at which at least one CCV formed direct IPs in primary luminal epithelial cells or fibroblasts. Number of CCVs – number of CCVs reported for the signal; Captured CCVs – number of CCVs that were targeted by the capture array. 'EPI' and 'FIB' columns show number of CCVs that formed direct IPs in the 2kb-binned luminal epithelial cells or fibroblasts, respectively. Fine-mapping region coordinates are in GRCh38/hg38. (*) – there were several fine-mapping regions (originally defined in hg19) that when lifted over to hg38 were split or partially deleted in hg38. These regions were compiled manually to encompass all CCVs at each of the regions. Red – a signal that was not targeted by the rCHi-C capture array, at which some of the CCVs (that mapped 'by chance' to the non-baited interacting bins) were involved in direct IPs.

However, at several signals, the rCHi-C data could not be used to prioritise a small subset of CCVs. For example, at chr8:126,412,414-129,029,685 (signal 2) there are 44 CCVs, of which 23 (EPI) and 38 (FIB) CCVs formed direct interaction peaks. This is partially due to the high CCV density at this signal – almost half of the CCVs span ~ 10 kb region, making it difficult to separate CCVs from each other even when the 2kb-binned data is used (EPI: 1 to 4 CCVs per bin; FIB: 1 to 8 CCVs per bin). As a result, higher resolution techniques and/or additional data types are required to prioritise CCVs at such signals.

5.6. Luminal epithelial cells versus fibroblasts

To investigate whether a subset of loci might mediate risk association via fibroblasts rather than epithelial cells, I compared 'direct' genes (and CCVs) identified between the two cell types (Figure 5.4).

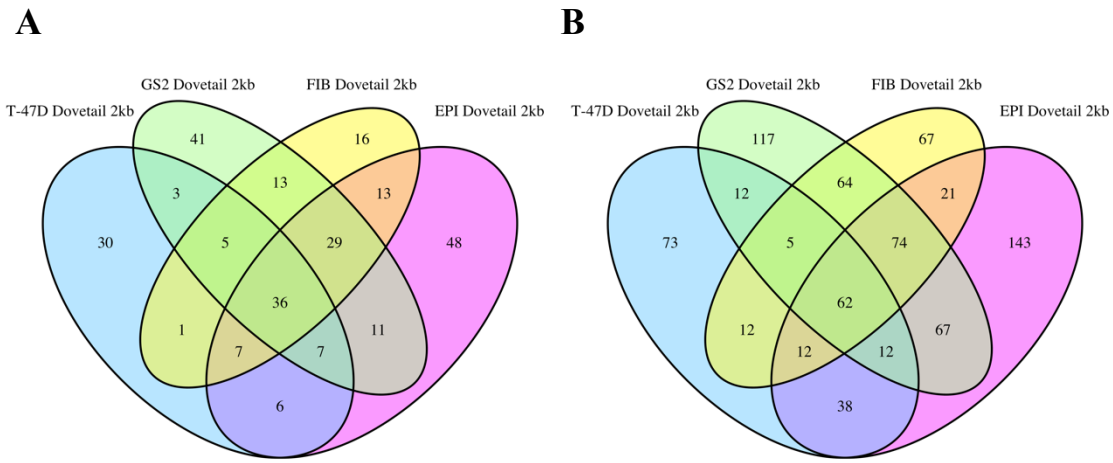

Figure 5.4: Venn diagrams illustrating the overlap between 'direct' genes and CCVs identified from cell line and primary cell rCHi-C libraries. 'Direct' genes (A) and 'direct' CCVs (B) identified in the 2kb-binned rCHi-C libraries generated using the Dovetail Genomics Omni-C protocol in T-47D, GS2, primary luminal epithelial cells (EPI) and primary fibroblasts (FIB).

Out of 157 and 120 unique 'direct' genes identified in the luminal epithelial and fibroblast libraries, respectively, 85 genes formed direct interaction peaks in both cell types, 72 genes only formed direct interaction peaks in the EPI dataset and 35 genes only formed direct interaction peaks in the FIB dataset ('direct' CCVs: in both – 169; EPI only – 260; FIB only – 148). Out of 35 genes that only formed direct interaction peaks in the FIB dataset, 16 participated in interaction peaks that did not involve CCVs (non-direct, hereafter) in the EPI dataset, and 19 genes formed no interaction peaks in the EPI dataset

at all. One such example is *FGF10* (Figure 5.5). The *FGF10* promoter was involved in four significant interaction peaks at the chr5:44,013,202-45,206,396 region (5p12 locus) in the FIB dataset, two of which were direct and the other two did not involve any CCVs. The direct interaction peaks involved two different CCVs – rs6885754 and rs199581089. rs6885754 also formed direct interaction peaks with the *NNT* and *PAIP1* promoters in the FIB dataset but no interaction peaks in the EPI dataset. The other CCV (rs199581089) also formed direct interaction peaks with *NNT* in both fibroblasts and epithelial cells, as well as it formed one interaction peak with each *FGF10* and *PAIP1* promoters in GS2 cells. According to the RNA-seq data (not shown), *FGF10* is expressed in both GS2 and primary fibroblasts, but not expressed in T-47D or primary luminal epithelial cells.

From Table 5.5, out of the 37 regions where at least one putative target gene was identified in both cell types, 12 regions were fully concordant, 23 were partially concordant, and 2 regions were completely different. These two regions are chr8:100,966,731-101,966,731 (8q22.3 locus) and chr10:21,244,013-22,620,463 (10p12.31). At the 8q22.3 locus (Figure 5.6), there is only one significant breast cancer risk signal that comprises 47 CCVs. In epithelial cells, a bin containing five closely spaced CCVs (rs35143639, rs13282693, rs544336840, rs36048804 and rs35794442) formed a direct interaction peak with the *GRHL2* promoter. In fibroblasts, another bin (~ 10 kb away) containing 6 different closely spaced CCVs (rs34113723, rs9297304, rs10086534, rs10086359, rs10089226 and rs16867595) formed a direct interaction peak with the *NCALD* promoter. *GRHL2* did not participate in any interaction peaks in fibroblast cells, while *NCALD* was involved in some non-direct interaction peaks in epithelial cells. Based on RNA-seq data (not shown), *GRHL2* is expressed at high levels in primary luminal epithelial cells but not expressed at detectable levels in primary fibroblasts. *NCALD*, in turn, is expressed in both primary fibroblasts and, at lower levels, in primary epithelial cells.

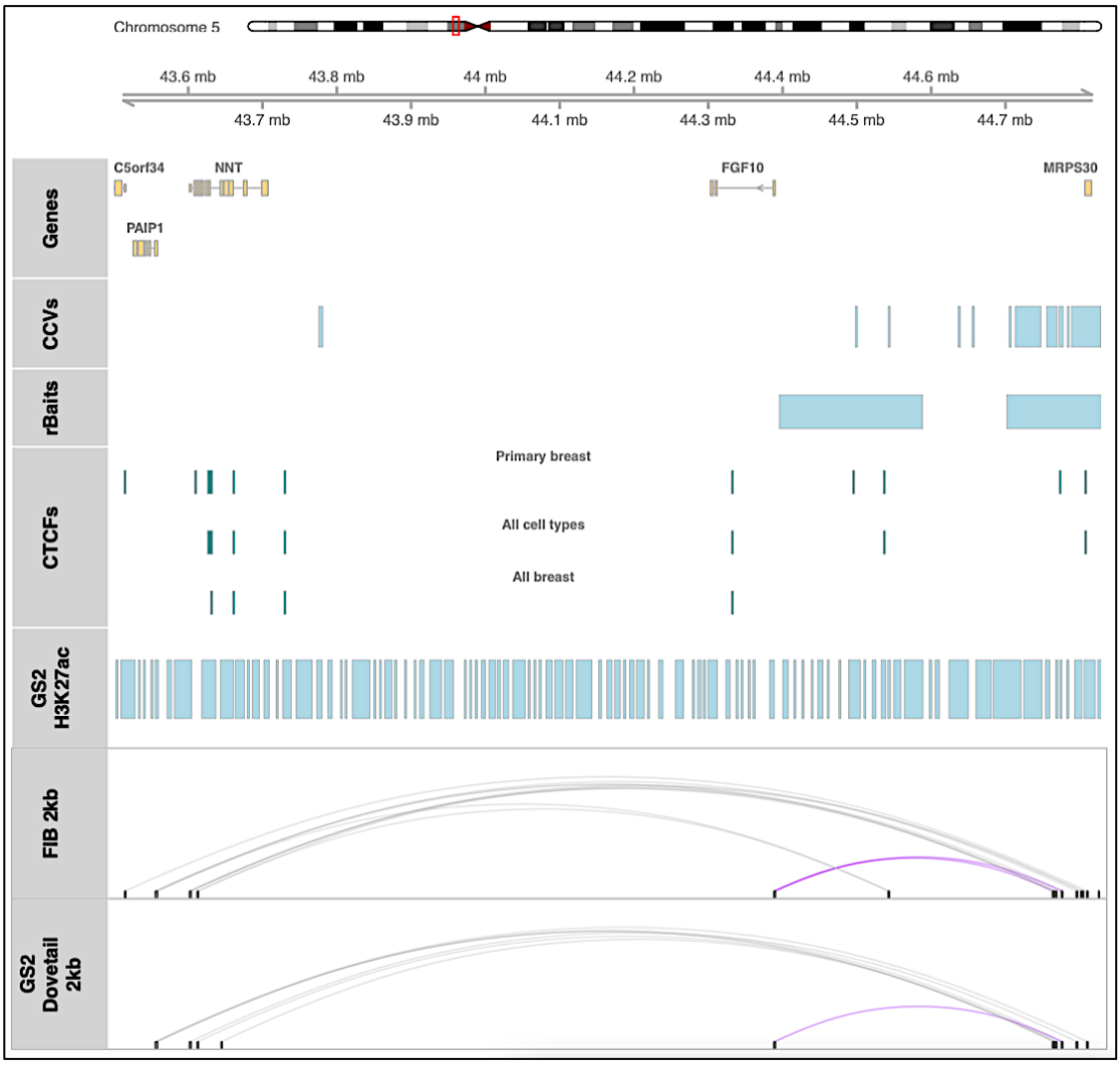

Figure 5.5: Direct interaction peaks at 5p12 in FIB and GS2 datasets. Direct interaction peaks (shown in a looping format) at the 5p12 breast cancer risk locus (chr5:44,013,202-45,206,396 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast fibroblasts (FIB) and GS2 cell line using the Dovetail Genomics protocol. Purple loops – direct IPs that involved the *FGF10* promoter. Grey loops – direct IPs that involved any other promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). GS2 H3K27ac – H3K27ac peaks identified from GS2 CUT&Tag data.

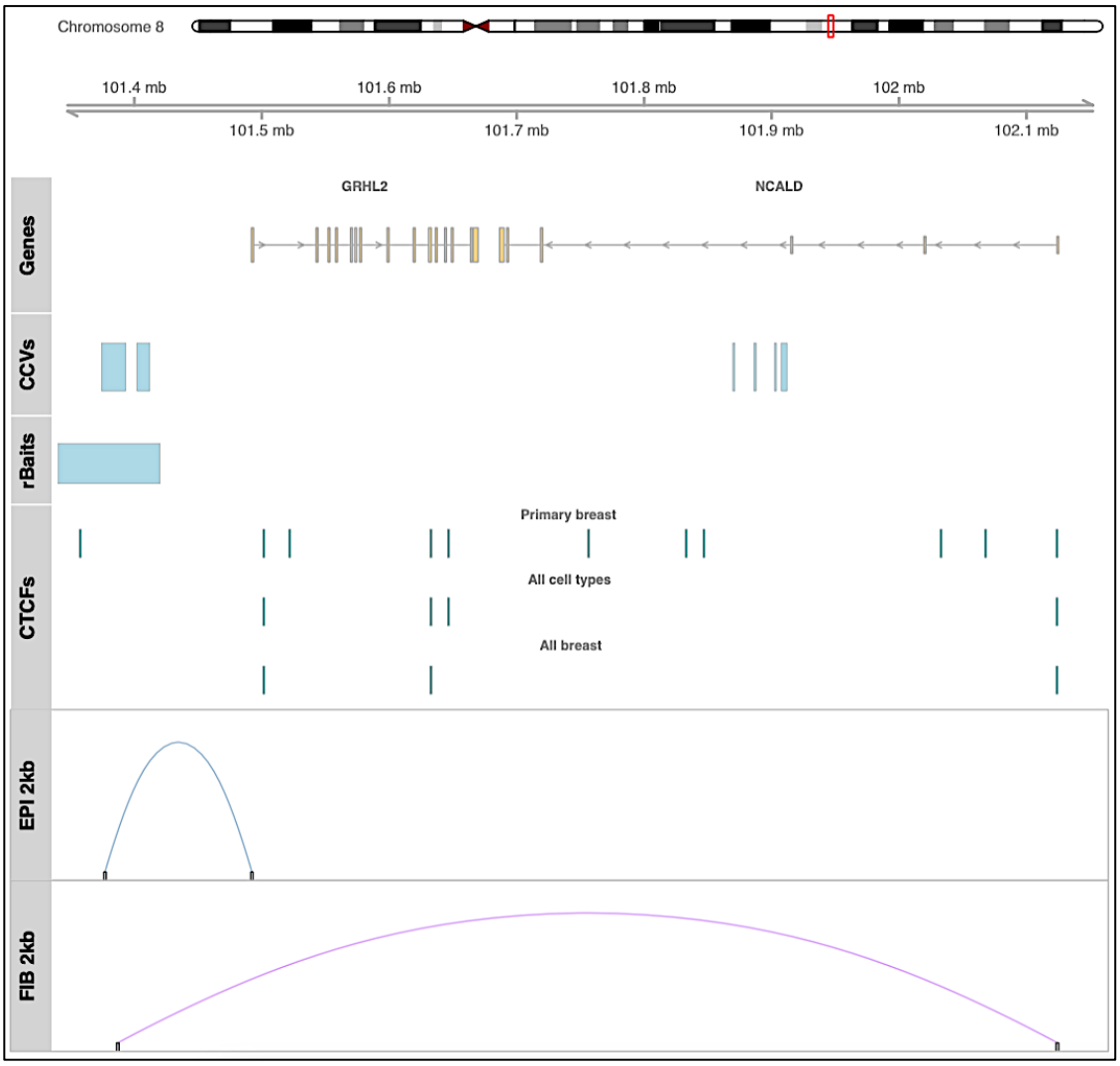

Figure 5.6: Direct interaction peaks at 8q22.3 in EPI and FIB datasets. Direct interaction peaks (shown in a looping format) at the 8q22.3 breast cancer risk locus (chr8:100,966,731-101,966,731 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast luminal epithelial cells (EPI) and fibroblasts (FIB) using the Dovetail Genomics protocol. Blue loops – direct IPs that involved the *GRHL2* promoter. Purple loops – direct IPs that involved the *NCALD* promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

At the 10p12.31 breast cancer risk locus (Figure 5.7), there are two independent 'strong-evidence' signals (signals 1 and 2). All 7 CCVs at signal 1 were covered by my capture array, while at signal 2, only 16 out of 58 CCVs were covered. In epithelial cells, rs10828247 (signal 1) formed a single direct interaction peak with the *BMI1* promoter. This CCV maps to the 5' untranslated region (UTR) of the *MLLT10* gene, so the CCV-containing bin also colocalised with the promoter of this gene. In fibroblasts, only rs56373249 (signal 2) formed a single direct interaction peak with the *MSRB2* promoter. *MSRB2* did not participate in any interaction peaks in epithelial cells.

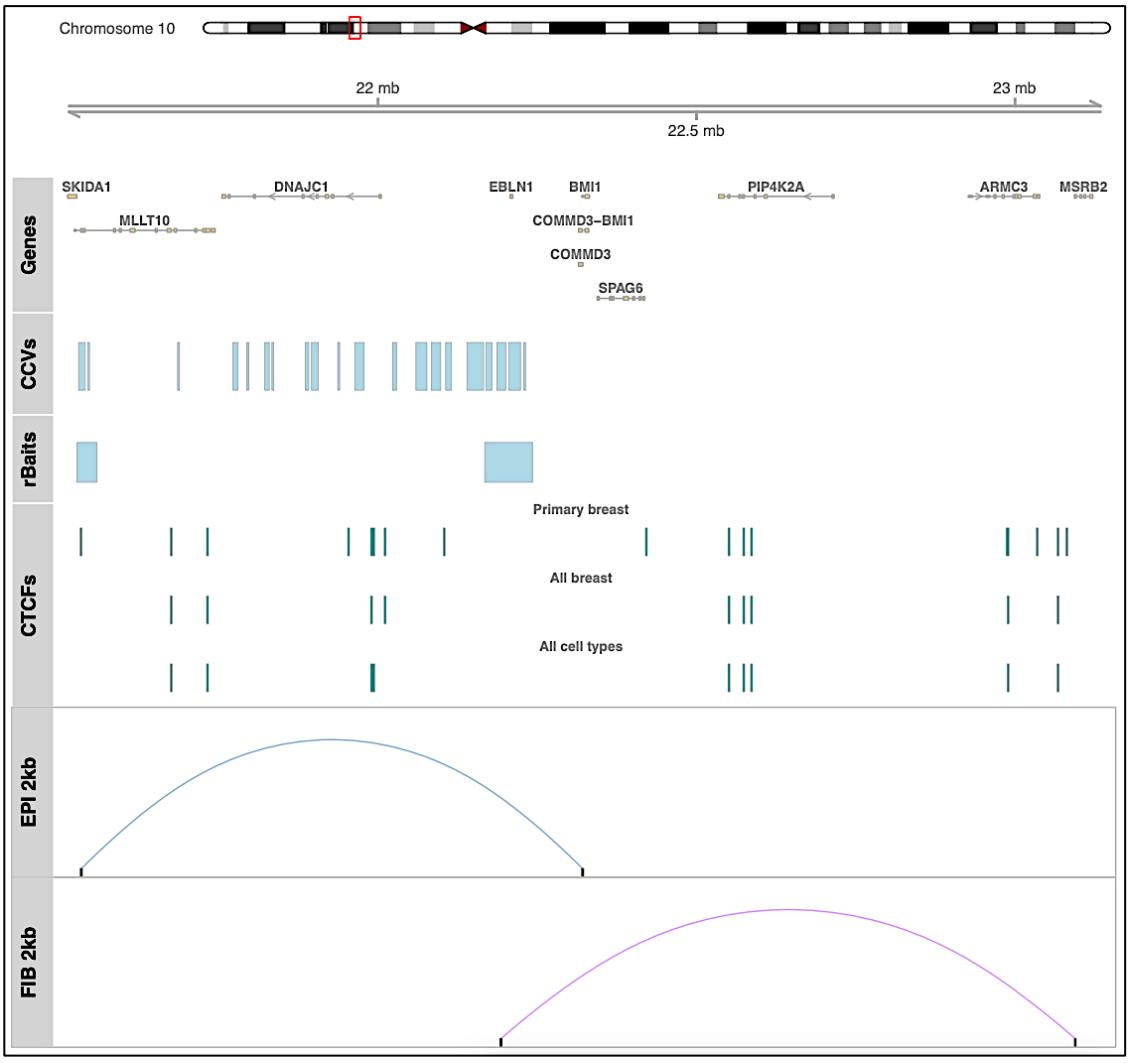

Figure 5.7: Direct interaction peaks at 10p12.31 in EPI and FIB datasets. Direct interaction peaks (shown in a looping format) at the 10p12.31 breast cancer risk locus (chr10:21,244,013-22,620,463 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast luminal epithelial cells (EPI) and fibroblasts (FIB) using the Dovetail Genomics protocol. Blue loops – direct IPs that involved the *BMI1* promoter. Purple loops – direct IPs that involved the *MSRB2* promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

Although at some signals EPI and FIB datasets supported the same or overlapping sets of CCVs (Table 5.6), there were several signals that looked profoundly different in the two cell types. One such example is at the chr6:129,527,974-130,527,974 region (6q22-q23 locus; Figure 5.8), that contains one signal (signal 1) encompassing 43 CCVs (all covered by my capture array). None of these CCVs formed direct interaction peaks in epithelial cells. However, 16 CCVs formed a total of 9 direct interaction peaks in fibroblasts. A bin containing two CCVs (rs4594967 and rs4551191) formed one direct interaction peak with each *AKAP7* and *EPB41L2* promoters. Two bins containing two (rs6900473 and rs6923819) and four (rs4404788, rs4279458, rs4321845 and rs75575024) CCVs formed

one direct interaction peak with each *SMLR1* and *EPB41L2* promoters. While three other bins containing one (rs9388766), three (rs9385532, rs7746589 and rs6914670) and four (rs7744830, rs7763108, rs11407151 and rs9375698) CCVs formed a single direct interaction peak each with the *EPB41L2* promoter only. Out of three genes, only *EPB41L2* was involved in non-direct interaction peaks in epithelial cells, while *AKAP7* and *SMLR1* formed no interaction peaks in the EPI dataset at all.

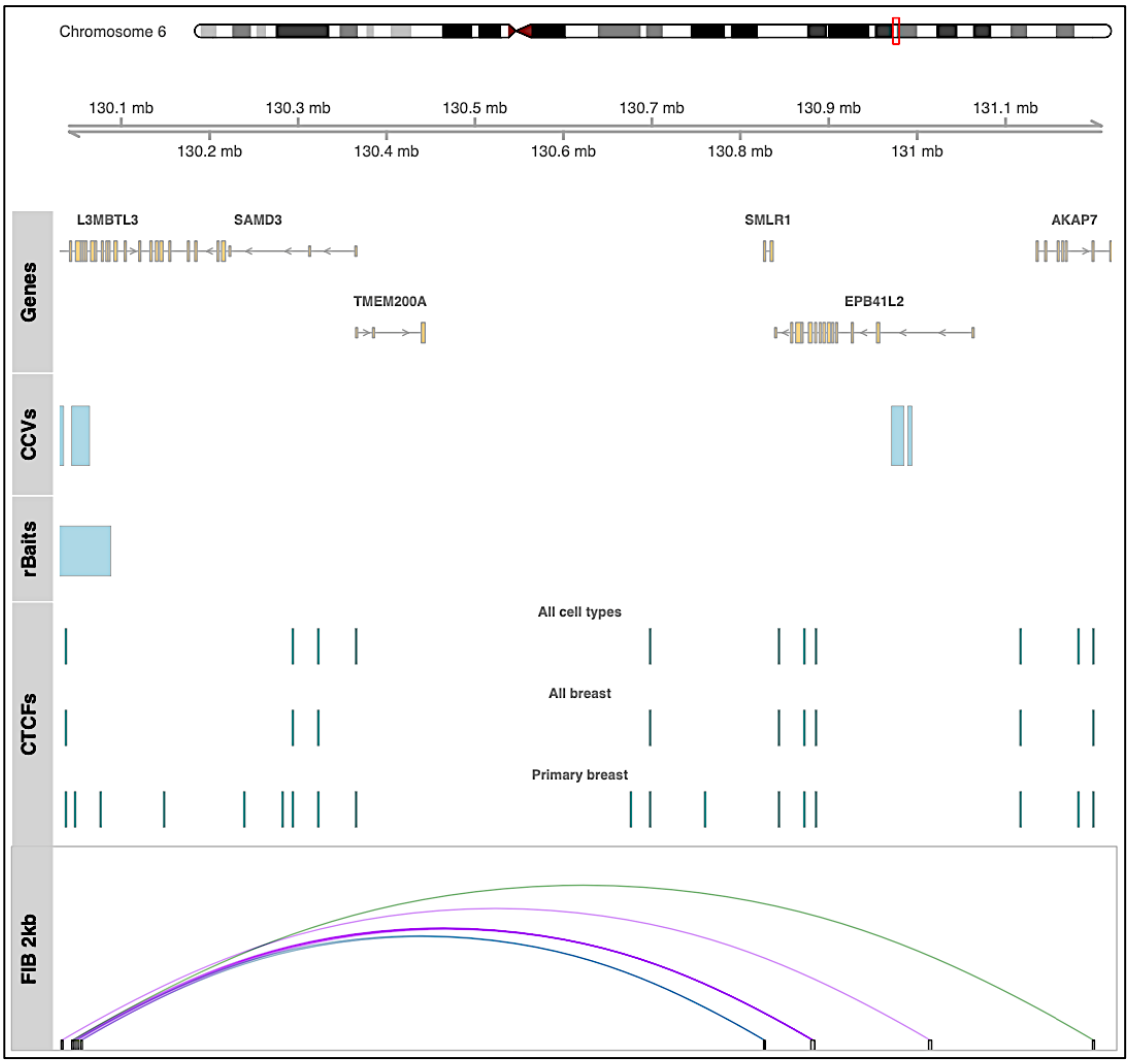

Figure 5.8: Direct interaction peaks at 6q22-q23 in FIB dataset. Direct interaction peaks (shown in a looping format) at the 6q22-q23 breast cancer risk locus (chr6:129,527,974-130,527,974 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast fibroblasts (FIB) using the Dovetail Genomics protocol. Blue loops – direct IPs that involved the *SMLR1* promoter. Purple loops – direct IPs that involved *EPB41L2* promoters. Green loops – direct IPs that involved the *AKAP7* promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

5.7. Primary cells versus cell lines

To evaluate cell lines as a model to recapitulate the disease, I compared 'direct' genes and CCVs identified in primary luminal epithelial cells and fibroblasts to those from T-47D and GS2 cell lines (Figure 5.4 and Table 5.7). Overall, higher overlap was observed between primary fibroblasts and GS2 cells rather than between primary luminal epithelial and T-47D cells.

Baseline cell type	Number of direct genes/CCVs	Comparison cell type	Direct genes/CCVs shared (%)
		Genes	
EPI	157	T-47D	56 (36%)
FIB	120	GS2	83 (69%)
EPI	157	FIB	85 (54%)
		CCVs	
EPI	429	T-47D	124 (29%)
FIB	317	GS2	205 (65%)
EPI	429	FIB	169 (53%)

Table 5.7: Comparison of 'direct' genes and CCVs identified in cell lines and primary cells.

I decided to focus on those genes (and CCVs) that appeared exclusively in:

- cells of the fibroblast lineage (Figure 5.4; GS2: 41 genes and 117 CCVs; FIB: 16 genes and 67 CCVs; GS2+FIB: 13 genes and 64 CCVs)

and

- cells of epithelial lineage (Figure 5.4; T-47D: 30 genes and 73 CCVs; EPI: 48 genes and 143 CCVs; T-47D+EPI: 6 genes and 38 CCVs).

Defining genes or CCVs that were exclusive to one library type as 'exclusive', I found that only a subset of 'exclusive' genes formed direct interaction peaks with a subset of 'exclusive' CCVs (GS2: 22 genes with 27 CCVs; FIB: 5 genes with 6 CCVs; T-47D: 19 genes with 30 CCVs; EPI: 22 genes with 44 CCVs). The remaining genes formed interaction peaks with CCVs that also interacted in other datasets, but with different genes. The remaining CCVs formed interaction peaks with genes that appeared in other datasets but in combination with different CCVs. This is in line with a single enhancer acting on multiple target genes, and a single gene being regulated by different cell type specific enhancers[109].

Comparing genes that formed direct interaction peaks exclusively in the EPI but not T-47D dataset, or vice versa, some of the EPI exclusive genes formed interaction peaks in T-47D data but with bins that lacked CCVs (hence, they did not count as direct interaction peaks), while others formed no interaction peaks in the T-47D data at all. The same was true for FIB and GS2 datasets. Two examples of genes that were only active in primary cells but not cell lines are *SKP1* (EPI) and *ITGA6* (FIB).

The *SKP1* promoter formed three direct interaction peaks at the chr5:132,571,366-133,571,367 region in epithelial cells (5q31.1 locus; Figure 5.9). These interaction peaks involved 3 different bins containing 2 CCVs (rs14355 and rs13718), 3 CCVs (rs7730930, rs3088225 and rs76836760) and 10 CCVs (rs571173399, rs56076449, rs56083805, rs60306856, rs62375248, rs67394705, rs76880525, rs72801470, rs62375249 and rs77509681). The bin containing 10 CCVs was the only one that formed a direct interaction peak with *SKP1* only, while the other two bins also formed a single direct interaction peak each with *C5orf15*. In addition, the bin containing three CCVs colocalised with the *HSPA4* promoter.

The *ITGA6* promoter formed a single direct interaction peak at the chr2:171,019,711-172,608,243 region (2q31.1 locus) with rs2356791 (Figure 5.10). The same CCV also formed a direct interaction peak with a bin containing *DLX2* and *DLX2-AS1*.

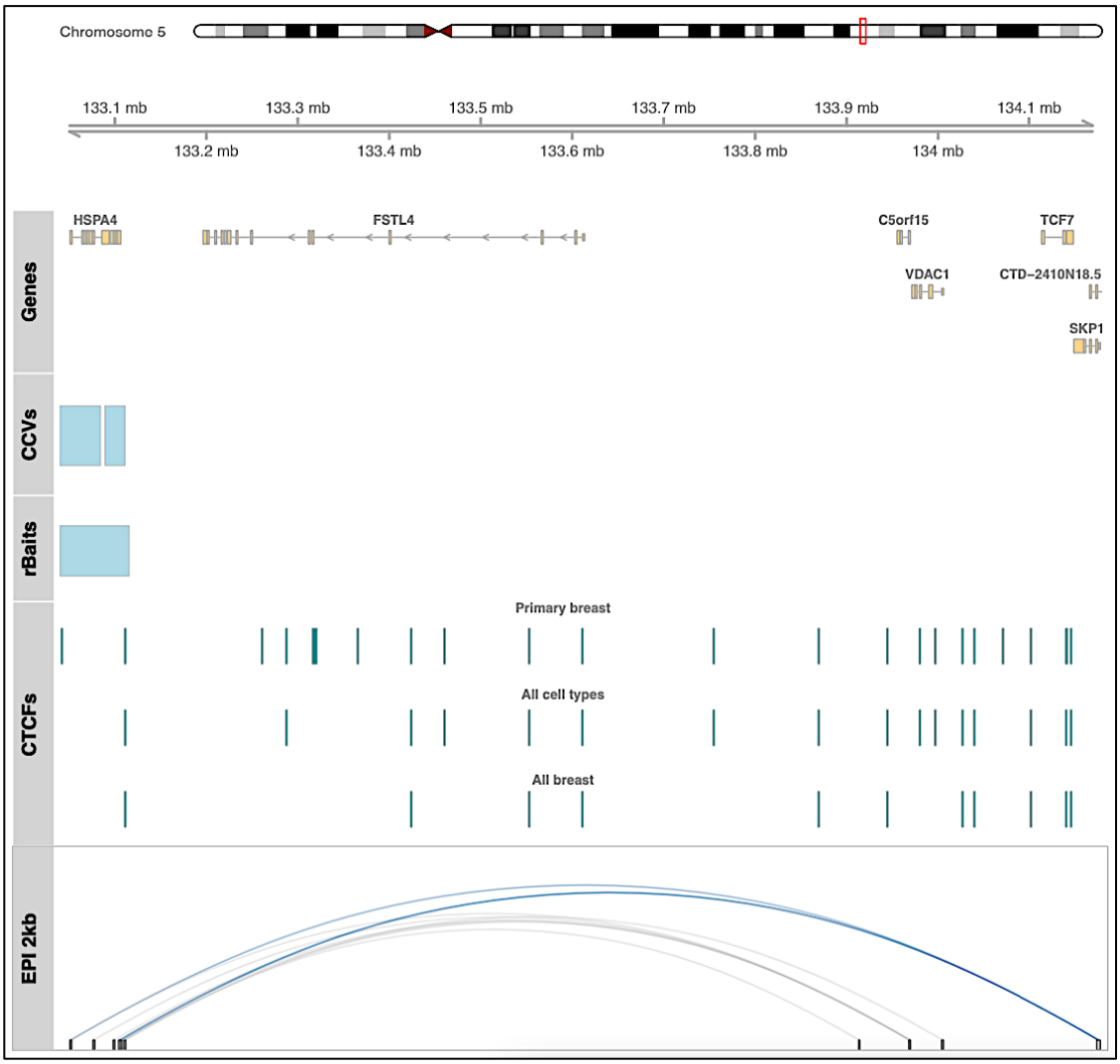

Figure 5.9: Direct interaction peaks at 5q31.1 in EPI dataset. Direct interaction peaks (shown in a looping format) at the 5q31.1 breast cancer risk locus (chr5:132,571,366-133,571,367 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast luminal epithelial cells (EPI) using the Dovetail Genomics protocol. Blue loops – direct IPs that involved the *SKP1* promoter. Grey loops – direct IPs that involved any other promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

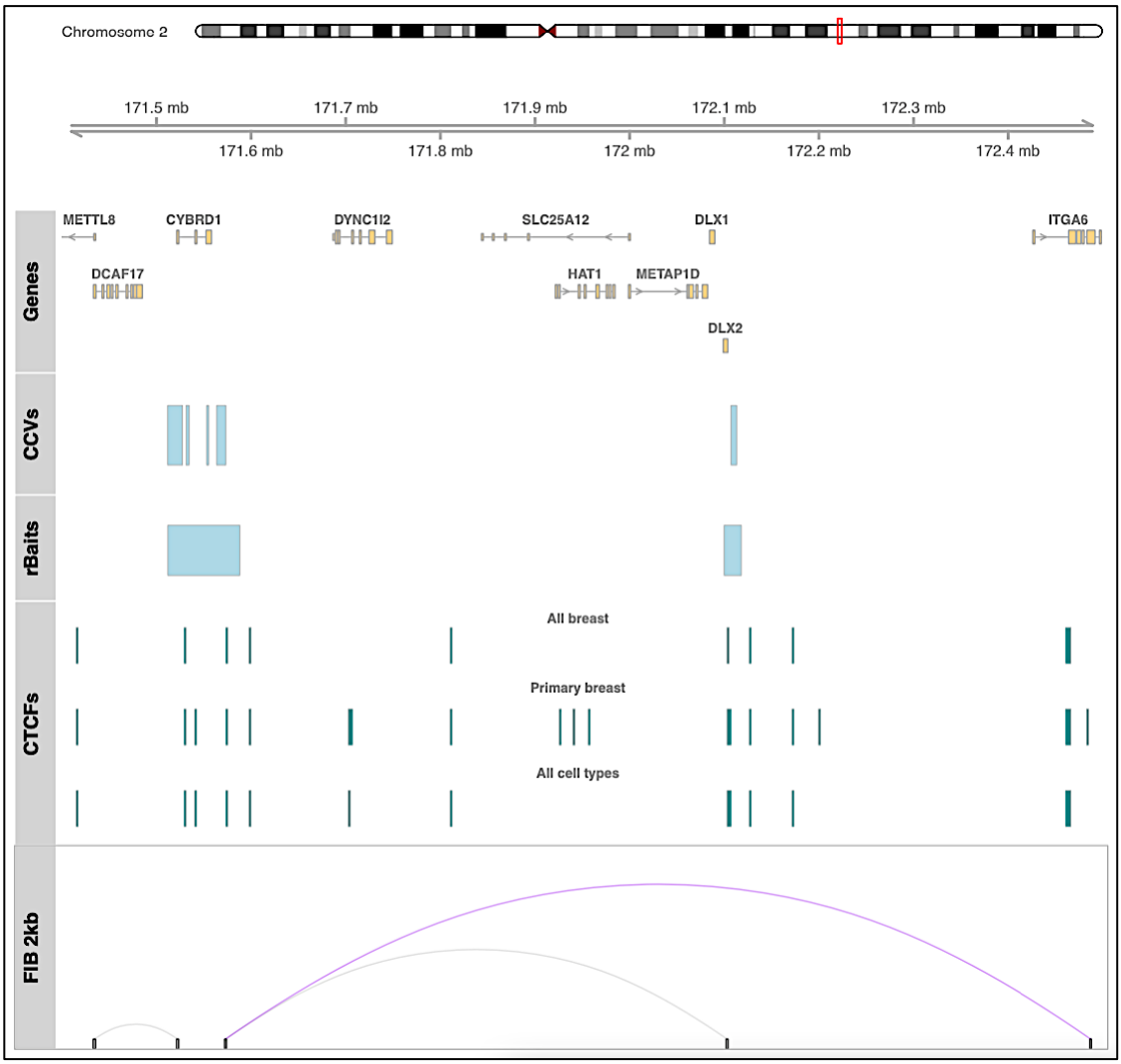

Figure 5.10: Direct interaction peaks at 2q31.1 in FIB dataset. Direct interaction peaks (shown in a looping format) at the 2q31.1 breast cancer risk locus (chr2:171,019,711-172,608,243 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast fibroblasts (FIB) using the Dovetail Genomics protocol. Purple loops – direct IPs that involved the *ITGA6* promoter. Grey loops – direct IPs that involved any other promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8).

Out of 28 regions where at least one putative target gene was identified in both EPI and T-47D datasets (Table 5.5 and Table 3.6), 8 regions were fully concordant, 16 were partially concordant, and 4 regions were completely different. One such example is signal 1 (17 CCVs) at the chr11:129,082,612-130,091,276 region (11q24.3 locus), where I identified *RP11-237N19.3* (T-47D) and *BARX2* (EPI) as the putative target genes (Figure 5.11). *BARX2* was involved in a single direct interaction peak with rs12285545 in luminal epithelial cells. In T-47D, rs139474311 formed a single bait-to-bait direct interaction peak with a bin located ~ 10 kb away and containing *RP11-237N19.3* and 3 other CCVs belonging to the same signal (rs11822830; rs11820646; rs11437753).

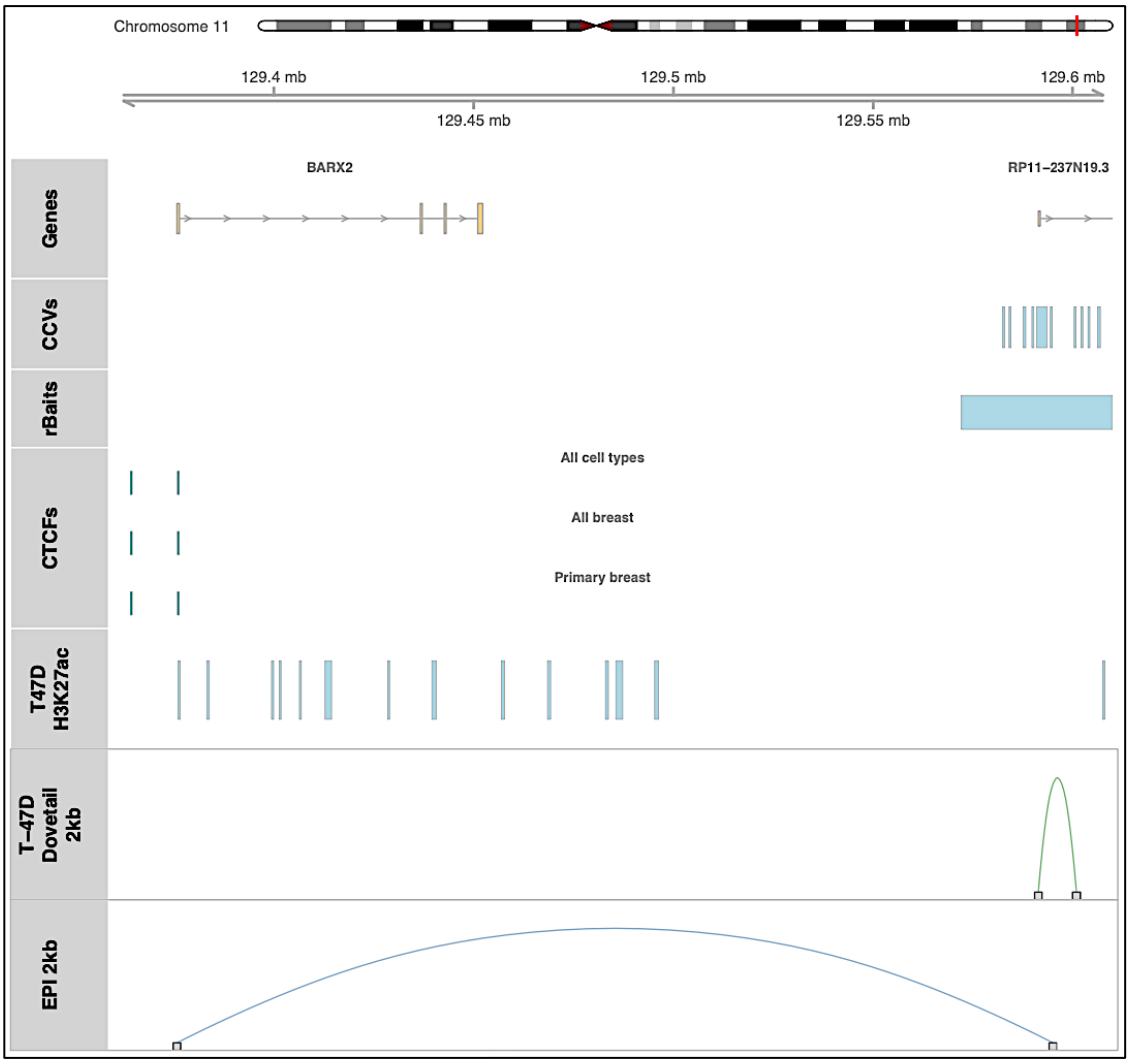

Figure 5.11: Direct interaction peaks at 11q24.3 in T-47D and EPI datasets. Direct interaction peaks (shown in a looping format) at the 11q24.3 breast cancer risk locus (chr11:129,082,612-130,091,276 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in T-47D cell line and in primary breast luminal epithelial cells (EPI) using the Dovetail Genomics protocol. Green loops – direct IPs that involved the *RP11-237N19.3* promoter. Blue loops – direct IPs that involved the *BARX2* promoter. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). T47D H3K27ac – H3K27ac peaks identified from T-47D CUT&Tag data.

Out of 41 regions where at least one putative target gene was identified in both FIB and GS2 datasets (Table 5.5 and Table 3.7), 15 regions were fully concordant, 23 were partially concordant, and 3 regions were completely different. An example here is signal 1 (2 CCVs) at the chr3:86,488,393-87,488,393 fine-mapping region (3p12-p11 locus), where rs13066793 formed a direct interaction peak with *LINC00506* in GS2 cells, and with a bin containing *CGGBP1* and *ZNF654* in fibroblast cells (Figure 5.12). Interestingly, the same interaction peak (between rs13066793 and *CGGBP1* and *ZNF654*) was also picked up in epithelial cells.

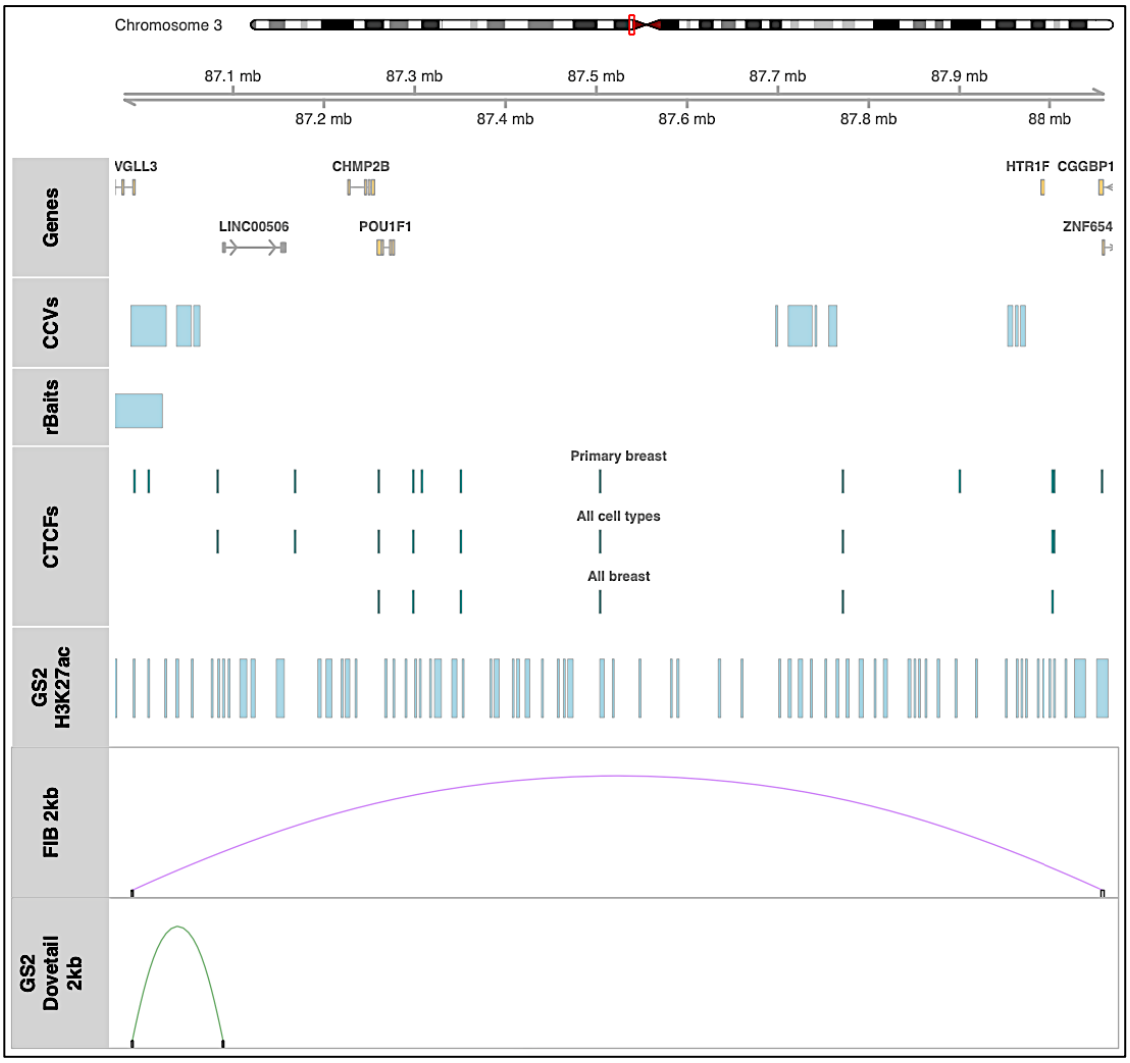

Figure 5.12: Direct interaction peaks at 3p12-p11 in FIB and GS2 datasets. Direct interaction peaks (shown in a looping format) at the 3p12-p11 breast cancer risk locus (chr3:86,488,393-87,488,393 fine-mapping region, hg38) detected in 2kb-binned rCHi-C data generated in primary breast fibroblasts (FIB) and GS2 cell line using the Dovetail Genomics protocol. Green loops – direct IPs that involved the *LINC00506* promoter. Purple loops – direct IPs that involved *CGGBP1* and *ZNF654* promoters. Genes – annotated RefSeq gene promoters. CCVs – credible causal variants selected by the BCAC fine-mapping study[67]. rBaits – rCHi-C array regions. CTCFs – consensus CTCF sites (described in Section 2.8). GS2 H3K27ac – H3K27ac peaks identified from GS2 CUT&Tag data.

Focussing on CCVs (Table 5.6 and Table 3.9), a similar pattern was observed, i.e., there were examples of concordance but also discrepancies between the cell line and primary cell data. For example, at the chr7:93,984,487-94,984,487 region (signal 1), where T-47D and GS2 data prioritised only 2 and 9 CCVs, 18 and 17 CCVs formed direct interaction peaks in EPI and FIB datasets, respectively. In contrast, at the chr8:100,966731-101,966,731 region (signal 1) where T-47D data prioritised 16 CCVs, rCHi-C data generated in epithelial cells prioritised a smaller number of CCVs (5 CCVs); and at the chr3:63,456,021-64,482,224 region (signal 1) where 12 CCVs formed direct interaction peaks in GS2 cells, only 2 did so in fibroblasts. In addition, at several signals where cell

line data prioritised a subset of CCVs, no CCVs formed direct interaction peaks in primary cells and vice versa. For instance, at the chr10:8,546,150-9,546,150 region (signal 1) no CCVs formed direct interaction peaks in GS2 data, while 6 CCVs formed direct interaction peaks in FIB data. At the chr17:79,794,855-79,816,335 region (signal 1), 6 CCVs formed direct interaction peaks in the T-47D dataset, while no CCVs did so in the EPI dataset. In addition, sometimes the same putative target gene formed direct interaction peaks with different CCVs (of the same signal) in primary cell compared to the cell line data. One of such examples is *ERLIN2* – a putative target gene of the chr8:36,500,965-37,501,668 region. In T-47D cells, *ERLIN2* formed a single direct interaction peak with rs56687477, while in primary luminal epithelial cells with rs10092900.

5.8. Discussion

So far CHi-C data have only been generated in breast cancer and immortalised 'normal' breast epithelial cell lines. Here, for the first time I generated CHi-C libraries in primary breast luminal epithelial cells and fibroblasts isolated from two women undergoing reduction mammoplasty. To increase the power, two biological replicates were pooled for the analysis. However, it would be informative to generate two technical replicates of each library to assess the similarities and differences observed between women.

Using rCHi-C data in primary cells, I identified a total of 157 genes that formed direct interaction peaks with 429 CCVs at 57 breast cancer risk fine-mapping regions in luminal epithelial cells and 120 genes interacting with 317 CCVs at 43 regions in fibroblasts. Without functional investigation, it is not possible to determine whether the identified genes are truly target genes and the CCVs are functional variants or not. In an attempt to address this in a high-throughput manner, the recent BCAC large-scale genetic fine-mapping analysis integrated *in silico* and functional genomic datasets into the analysis using their INQUISIT algorithm[66]. Comparing my lists of putative target genes with the 'high confidence' INQUISIT predictions, I found a greater concordance between the INQUISIT target gene predictions and the direct genes predicted from the EPI data compared to those predicted from the FIB data (Table 5.5). This may reflect the fact that epithelial cells are more relevant, because it is an epithelial cell (a luminal progenitor cell) that is generally considered to be the cell of origin for breast cancer[105]. Alternatively, it may reflect the fact that the data types that are incorporated into the INQUISIT algorithm

are predominantly generated in (epithelial) breast cancer cell lines. Overall, quite a low proportion of INQUISIT 'high confidence' genes were identified as direct genes using primary cell rCHi-C data (~20% in EPI; ~12% in FIB). One possible explanation is the lack of cell type diversity in the genomic features that are incorporated into the INQUISIT: out of the 811 genomic features incorporated into the algorithm, 362 (44.6%) were generated in MCF-7 cells (an ER+ breast cancer cell line), and just 67 (8.3%) or 22 (2.7%) were generated in primary mammary epithelial or luminal epithelial cells, respectively[130].

To guide the interpretation of regulatory variants, tools for the annotation of CCVs have been developed (such as HaploReg[131] and RegulomeDB[132]). These tools align CCVs with markers of open chromatin (such as DNase-seq and ATAC-seq), active histone modifications (particularly H3K27ac, H3K4me1 and H3K4me3) and TF-binding sites generated in different cell types. However, the range of assays and cell types used in these tools is limited. Both HaploReg and RegulomeDB primarily use data from the ENCODE[109] and Roadmap Epigenomics project[133]. ENCODE relies heavily on cell lines (MCF-7, MCF10A and T-47D), with 267 out of a total 468 datasets generated in MCF-7 cells[130]. Although the Roadmap Epigenomics project uses primary *ex vivo* tissues to generate normal epigenomes (which are arguably more relevant for analyses of breast cancer risk, given that risk reflects early events that precede the somatic genome), the range of data types is, inevitably, more limited, and with most datasets generated in myoepithelial cells[130]. Due to these limitations, functional validation of the variants is required, such as high-throughput reporter gene assays (e.g., MPRA[134] and STARR-seq[135]), but these have not yet been used in the context of breast cancer GWAS.

To assess the level to which rCHi-C data can help to prioritise risk-associated variants at GWAS risk loci, 'direct' CCVs were mapped back to the breast cancer risk signals. At most of the signals where my capture array covered all reported CCVs and where at least one CCV formed direct interaction peaks with at least one gene, interaction peaks in the rCHi-C data involved only a subset of the CCVs, potentially narrowing down the number that would be prioritised for the follow up studies (Table 5.6).

For example, out of 13 CCVs at the 2q31.1 breast cancer risk locus (Figure 5.2), I prioritised a single variant (rs930313) that formed direct interaction peaks with *OLA1* in EPI and with *LINC01305* in EPI and FIB. According to RegulomeDB and HaploReg,

rs930313 had the highest probability score of 0.61 (together with four other CCVs) and mapped to multiple marks of open chromatin in the largest number of tissues, implying that it might be a functional variant. Long non-coding RNA *LINC01305* has been recently linked to the cervical cancer progression[136, 137], while *OLA1* was demonstrated to promote tumour invasion and metastasis in breast cancer by inhibiting the production of reactive oxygen species[138], and to enhance chemoresistance by inhibiting the epithelial-mesenchymal transition (EMT) process in breast cancer cells[139].

I also prioritised a subset of CCVs at the 1q32.1 breast cancer risk locus (Figure 5.3), which formed direct interaction peaks with *ETNK2* and *SOX13* promoters in both primary cell types. Out of seven prioritised CCVs, rs7520079, an intron variant of *SNRPE* gene, has the highest RegulomeDB probability score of 1.0 and is predicted to colocalise and alter the consensus binding motif of an E3 ubiquitin ligase (TOPORS) that has been implicated in modulating sensitivity to a PARP inhibitor olaparib in androgen receptor positive breast cancer cells[140]. Although analysis of TCGA samples revealed that *ETNK2* is amplified in 13% of breast cancer patients[141], its function in breast cancer remains unknown. However, a recent study proposed its role in gastric cancer[142]. It showed that *ETNK2* was upregulated in patients with hepatic metastasis. *ETNK2* knockout significantly reduced proliferation, invasion, and migration and increased apoptosis. In mouse xenograft models, *ETNK2* knockout virtually abolished hepatic metastasis. Studies suggest emerging role of SOX proteins in breast cancer development and maintenance[143]. *SOX13* specifically was found to be upregulated in breast cancer tissues and cells compared with normal samples. Knockdown of *SOX13* inhibited breast cancer cell proliferation, arrested cell cycle at G1/S phase and suppressed glycolysis, while overexpression of *SOX13* reversed these events[144].

However, at the 'dense' regions where multiple correlated variants all map closely to each other, CHi-C was not very effective in reducing the number of candidate CCVs. As a result, higher resolution techniques and/or additional data types are required to prioritise CCVs at such signals.

5.8.1. Luminal epithelial cells versus fibroblasts

To investigate whether a subset of loci might mediate risk association via fibroblasts rather than epithelial cells, I compared 'direct' genes (and CCVs) identified between the

two cell types (Figure 5.4). A large proportion of genes formed direct interaction peaks in both cell types, but there were 35 genes that did so exclusively in fibroblasts. To illustrate this, I picked *FGF10* at the 5p12 locus (Figure 5.5). *FGF10* is expressed in primary fibroblasts and GS2 cells (where it also formed direct interaction peaks), but not in primary luminal epithelial or T-47D cells. The fact that there were two other genes (*NNT* and *PAIP1*) that formed direct interaction peaks with the same CCVs at that region does not rule out *FGF10* as a plausible putative target gene at this locus that might as well be involved in breast cancer risk association. Studies in FGF10−/− and FGFR2b−/− mouse embryos have shown that FGF10-FGFR2b signalling plays a key role in mammary gland development[145, 146]. *FGF10* was found to be highly overexpressed in 10% of human breast carcinomas[147]. The gene is expressed exclusively by the stromal fibroblasts of normal and breast cancer tissue and has been reported to be an oncogene in mammary tumour virus mouse models and in a subset of breast carcinomas where it is overexpressed[147, 148]. *FGF10* is involved in regulation of the EMT, cell viability, migration and colony formation in breast cancer cell lines[149]. *FGF10* has previously been proposed as a target gene at the 5p12 breast cancer risk locus, on the basis of expression quantitative trait locus analysis (eQTL) in normal breast tissues and breast tumours, as well as 3C and reporter gene assays carried out in cell lines of epithelial origin[150]. My rCHi-C data, however, implicated *FGF10* as a putative target gene that acts in fibroblasts, but the direct interaction peaks I picked up were with signal 3 CCVs, while Ghoussaini et al. study[150] focused on signal 1 CCV.

To illustrate further potential cell type-specific differences in my data, I used the 8q22.3 locus (Figure 5.6) at which *GRHL2* and *NCALD* formed direct interaction peaks in EPI and FIB, respectively; and the 10p12.31 locus (Figure 5.7) at which *BMI1* formed direct interaction peaks in EPI and *MSRB2* in FIB. *GRHL2* is a transcription factor that is suggested to play an important role in EMT[151]. Knockdown of *GRHL2* expression in human mammary epithelial cells led to down-regulation of E-cadherin and induction of EMT. Clinical datasets showed that expression of *GRHL2* is associated with poor relapse free survival and increased risk of metastasis in breast cancer patients. In mouse models, overexpression of the gene significantly promoted tumour growth and metastasis. *NCALD* showed lower expression in the stroma surrounding invasive breast primary tumours than in normal breast stroma cells[152]. Studies showed that *NCALD* is involved in chemoresistance in ovarian[153] and colorectal[154] cancers. In colorectal cancer, miR-181d-5p, microRNA (miRNA) from cancer-associated fibroblasts (CAFs)-secreted

exosomes, directly targeted *NCALD* to inhibit the 5-Fluorouracil-based chemotherapy sensitivity of colorectal cancer cells. Patients with higher *NCALD* levels showed a higher survival rate. Interestingly, CAF-secreted exosomes containing miR-181d-5p induced EMT in breast cancer cells and promoted tumour growth in mouse models[155].

At the 10p12.31, *BMI1* and *MSRB2* formed direct interaction peaks with rs10828247 (signal 1) and rs56373249 (signal 2), respectively, suggesting that either or both genes might represent targets albeit of different signals. Transcriptional repressor *BMI1* is an established oncogene that was linked to multiple cancers[156]. *MSRB2* is a lot less studied, especially in the context of cancer. Downregulation of its family gene *MSRA* in human breast cancer cells resulted in increased cell proliferation and extracellular matrix degradation, and consequently in a more aggressive cellular phenotype[157]. Another family gene *MSRB3* was demonstrated to prevent oncogene-induced DNA damage in breast cancer[158]. *MSRB2* itself is suggested to play a role in the induction of mitophagy – a selective degradation of toxic mitochondria that protects a cell from apoptosis[159]. Senescent human fibroblasts showed decreased expression of *MSRB2* when compared to young cells, suggesting that gene downregulation may alter the ability of senescent cells to cope with oxidative stress and result in the age-related accumulation of oxidative damage[160]. Analysis of primary cancer samples showed that *MSRB2* is highly expressed in liver and papillary renal cancers[161].

Although functional studies are required, it is possible that some of the genes mentioned in this sub-chapter may represent cell type specific targets, some of which may be involved in mediating breast cancer risk via fibroblasts.

5.8.2. Cell lines as model systems

Cell lines are *in vitro* model systems that are widely used in cancer research. The main advantage of cell lines lies in their ability to provide an indefinite source of biological material for experimental purposes. However, despite being a powerful tool, cell lines are genetically manipulated which can affect their phenotype, native functions and responsiveness to stimuli. Serial passage of cell lines can lead to further genotypic and phenotypic variations over time, as well as heterogeneity within cultures at a single point in time. As a result, cell lines may not adequately represent primary cells for certain purposes.

To evaluate T-47D and GS2 cell lines as model systems for studying the mechanisms that drive breast cancer risk in epithelial cells (T-47D) and fibroblast (GS2), I compared 'direct' genes and CCVs identified in primary luminal epithelial cells and fibroblasts to those from T-47D and GS2 cell lines (Figure 5.4 and Table 5.7). Overall, using the primary cell data as the baseline for comparison, the proportion of shared 'direct' genes in the EPI and T-47D datasets (36%) was much lower than the proportion of shared genes in the FIB and GS2 data sets (69%; $p = 5.3 \times 10^{-8}$). In fact, the proportion of shared 'direct' genes between EPI and FIB (54%) was higher than between EPI and T-47D (36%, $p = 0.001$). The opposite was true comparing EPI and FIB (54%) to FIB and GS2 (69%, $p = 0.01$). The same trends were observed when comparing the proportions of shared CCVs; however, these, in particular, need to be interpreted with caution, since they are not independent observations (i.e., multiple CCVs can cluster within bins). These comparisons suggest that GS2 may be a better model for mammary fibroblasts than T-47D is for normal mammary luminal epithelial cells. This may reflect the fact that GS2 is a 'normal' immortalised breast fibroblast cell line, while T-47D is a breast cancer cell line (not a 'normal' immortalised luminal epithelial cell line). However, these differences may also reflect the quality of T-47D libraries, since the data generated in T-47D cells tended to be of a poorer quality than the data generated in other cell types.

I selected *SKP1* (EPI) and *ITGA6* (FIB) as examples of genes that formed direct interaction peaks in primary cells but not cell lines (Figure 5.9 and Figure 5.10). *SKP1* is one component of the SCF E3 ubiquitin ligases that comprise three invariable components (SKP1, Cullin1, and RBX1) and a variable component – F-box proteins that determine substrate specificity (such as *FBXO32*, for example, that was also a putative target gene in luminal epithelial cells). A study showed that *SKP1* regulates BRCA1 protein stability[162]. In addition, *SKP1* expression was found to be significantly reduced (~25-fold) within invasive breast carcinomas compared to the normal tissues, and analysis of gene copy number alterations revealed that *SKP1* alterations occur in over 40% of breast cancer samples[163].

A transmembrane glycoprotein adhesion receptor protein *ITGA6* is abnormally expressed in multiple tumour types, including breast cancer. In addition to mediating interactions with the extracellular matrix, integrins drive intracellular signalling events that communicate from the tumour microenvironment to inside of the tumour cell to alter phenotypes including migration and invasion[164]. It has also been demonstrated that

ITGA6 is a hypoxia-inducible factors dependent target gene, and that its high expression enhances invasion and tumour-initiating cell activities in metastatic breast cancer models. *ITGA6* has also been proposed to be an independent prognostic factor for survival in breast cancer patients. ITGA6/AKT/ERK signalling is suggested to play an important role in radiotherapy resistance in human breast cancer[165]. In addition, a study of the breast cancer tumour microenvironment revealed that *ITGA6* was unregulated in the dense fibrotic zone of IDC[166]. Fibrosis is the formation of excess fibrous connective tissues due to physiological stress; when cancer becomes invasive or metastatic, dense fibrosis is detected around the tumour burden, especially in solid tumour tissue. Finally, studies showed that *ITGA6* renders an invasive phenotype on fibroblasts[167], and its knockdown significantly attenuates the proliferation and differentiation of fibroblasts into myofibroblasts[168].

Similarly, there is biological plausibility in the example that I used to illustrate a locus (11q24.3; Figure 5.11) at which there were different direct interaction peaks in primary luminal epithelial cells (with rs12285545 and *BARX2*) and T-47D (with rs139474311 and *RP11-237N19.3*). According to RegulomeDB, rs12285545 has a much higher probability score than rs139474311 (0.61 vs. 0.11). In addition, a HaploReg search showed that rs12285545 overlapped enhancer histone marks in four tissues (including breast myoepithelial primary cells). Not much information is available on *RP11-237N19.3*. *BARX2*, in turn, was shown to be involved in *ESR1* regulation[169]. Its protein binds to an *ESR1* gene promoter and increases the expression of alternatively spliced mRNAs that encode two ESR1 protein isoforms. Additionally, *BARX2* increases the expression of active matrix *MMP9*, which is known to promote invasion of cancer cells via matrix degradation.

Overall, these data suggest that although cell lines are useful model systems, they might not always accurately replicate the primary cells. If the risk reflects early events that occur in (relatively) normal cells, then further studies in primary cells are needed and, where it is not possible to use primary cells immortalised 'normal' mammary epithelial cell lines (such as MCF10A or Bre80) might be better models than the frequently used breast cancer cell lines (T-47D and MCF-7).

6. Discussion

Genome-wide association studies together with fine-mapping and large-scale replication studies have identified genetic variants associated with breast cancer risk in over 150 genomic regions[67]. However, the causal variants and targets genes that drive these associations remain largely unknown, with less than 20 regions studied in detail (Table 1.1). Although individual regions tend to account for a relatively modest proportion of risk association, when considered together, they might aid in selecting the individuals who are at high risk of developing breast cancer within a population.

Most of the identified CCVs map to non-protein-coding regions of the genome and are thought to affect transcriptional regulation[78-80]; many are found in gene deserts with the nearest known protein-coding genes mapping hundreds of kilobases away. It has been proposed that transcriptional regulation involves direct physical interaction between the regulatory element and target(s) which is brought about by chromatin looping[170]. Regulatory elements can be located a long distance away from their target genes and do not necessarily regulate the closest promoter, but evidence suggests that most enhancer-promoter interactions occur in *cis*[111, 116, 171]. In addition, studies have shown that individual enhancers can loop to and regulate multiple genes and that individual genes can be regulated by multiple enhancers[109].

Although GWAS have proven to be a powerful tool to identify disease-associated genetic variants, they do not directly address the underlying biological mechanisms. In addition, at many risk regions local correlation of multiple genetic variants due to LD is such that it makes it difficult to distinguish causal variants from a large number of correlated variants. Consequently, additional studies are generally required to identify and/or prioritise a subset of putative causal variants and target genes for in-depth functional characterisation.

Here I used CHi-C – a chromosome conformation capture-based method that allows high-throughput and high-resolution analysis of physical interactions between regulatory elements and their target genes. Until recently, the cellular input requirements have prohibited the use of CHi-C in all but a very few primary cell types[172] and the use of 6-cutter restriction enzymes (such as HindIII) has limited the resolution of the technique. In this project, I compared three different CHi-C protocols and found that by switching

to kit-based methods, I was able to reduce the cellular input, time and costs associated with the technique. I was also able to increase the resolution, which, in turn, allowed me to narrow down the number of prioritised genes and CCVs. In summary, I have generated CHi-C libraries in two types of primary cells (luminal epithelial cells and fibroblasts) and analysed the data at a resolution of 2 kb.

An advantage of CHi-C over many other methods used for the functional annotation of GWAS risk loci is that it links CCVs and putative target genes. While there are some other methods that link CCVs to genes (such as HiChIP[173] and CROP-seq[174]), most methods tend to focus on investigation of genes or variants separately. Data presented in this thesis suggest that CHi-C has a role to play in the functional annotation of GWAS risk loci, prioritising putative target genes and CCVs for functional follow up studies. However, integrating CHi-C data with other genomics datasets will be needed to further inform our understanding of the mechanisms that influence risk. For example, other studies have used CHi-C data combined with ChIP-seq and RNA-seq to investigate the rewiring of promoter–enhancer contacts upon the differentiation of embryonic stem cells[175-177]. In some of the examples that I have used to illustrate a point, I have referred to H3K27ac CUT&Tag and RNA-seq data generated in the same cell types by other members of the lab. As a group, we are currently integrating these datasets which will allow us to determine whether the specific examples I have mentioned in this thesis are representative of general trends.

In this project, I also investigated the benefit of integrating two CHi-C approaches (rCHi-C and pCHi-C). Looking from two opposite capture viewpoints allowed me to look into a subset of reciprocal interaction peaks, cross-check lists of prioritised genes and CCVs, and to investigate a subset of 'indirect' (third-party) interactions peaks. My study suggests that a dual-capture approach can aid in validation of interaction peaks, putative target genes and CCVs identified by one approach as well as in identification of additional targets. However, it also highlighted how important experimental design and analysis considerations are required in order to get the most out of the integration analysis.

6.1. Data analysis considerations

Capture Hi-C data has some challenging statistical properties that differentiate it from the Hi-C data. For instance, in CHi-C, the interactions are asymmetric because the number

of captured viewpoints is far less than the number of potential interacting non-baited ends. Additionally, baited regions may be captured with different efficiencies, affecting the background of the experiment. As a result, specialised tools are required for the analysis of the CHi-C data. Here I used CHiCANE[108] – an in-house pipeline developed specifically for the analysis of rCHi-C data. CHiCANE was designed on the assumption that GWAS risk loci harbour causal variants, some of which influence gene expression via long-range interactions with their target genes and prioritises mid-range interaction peaks (100 kb – 5 Mb).

Standard libraries were called using individual HindIII fragments as the unit of analysis, while Arima and Dovetail libraries were called using both 2kb- and 5kb-binned data. In both the cell line and primary cell rCHi-C data, more interaction peaks were called in the 2kb- compared to 5kb-binned data. In addition, higher proportions of *cis* interaction peaks in the ≥ 1 Mb range were consistently observed in the 5kb-binned (and Standard) libraries than in the 2kb-binned libraries.

In pCHi-C data, in contrast, more (or similar numbers) of interaction peaks were called in the 5-kb binned datasets. In addition, the 5kb-binned datasets showed greater proportions of *cis* interaction peaks in the 100 kb – 1 Mb range and lower proportions of *cis* interaction peaks in the 10 kb – 100 kb range. Comparing back to the rCHi-C libraries, the overall proportions of *cis* ≥ 1 Mb interaction peaks were generally lower in pCHi-C than in rCHi-C libraries.

There are at least two possible explanations for these differences. First, the pCHi-C analysis was based on a much larger number of on-target di-tags than the rCHi-C analysis (even though less of these di-tags mapped to the vicinity of a GWAS region). However, down-sampling the pCHi-C data to a similar size did not alter the distribution of called interaction peaks substantially (not shown). Alternatively, in the pCHi-C data, the baited smart bins are, on average, 2 kb in size (median = 1.9 kb), while the majority of non-baited bins are either 5 kb or 2 kb (Figure 6.1). In the rCHi-C data, the baits were selected based on the regions of LD (rather than gene promoters) and, therefore, most of them are not smart bins. As a result, both baited and non-baited bins in the rCHi-C datasets are predominantly 5 kb (or 2 kb) in size. The fact that the distribution of the called interaction peaks was more similar for the rCHi-C and pCHi-C datasets in the 2kb-binned analyses rather than the 5kb-binned analyses suggests that the size of the baited and non-baited

bins influences the predominant distance range within which interaction peaks are called, with smaller bin sizes favouring shorter-range interaction peaks, and with the size of the baited bins having the more pronounced effect. This observation uncovers another potential advantage of kit-based protocols (where data can be binned in various bin sizes) over the Standard protocol (where the size of the bin is fixed).

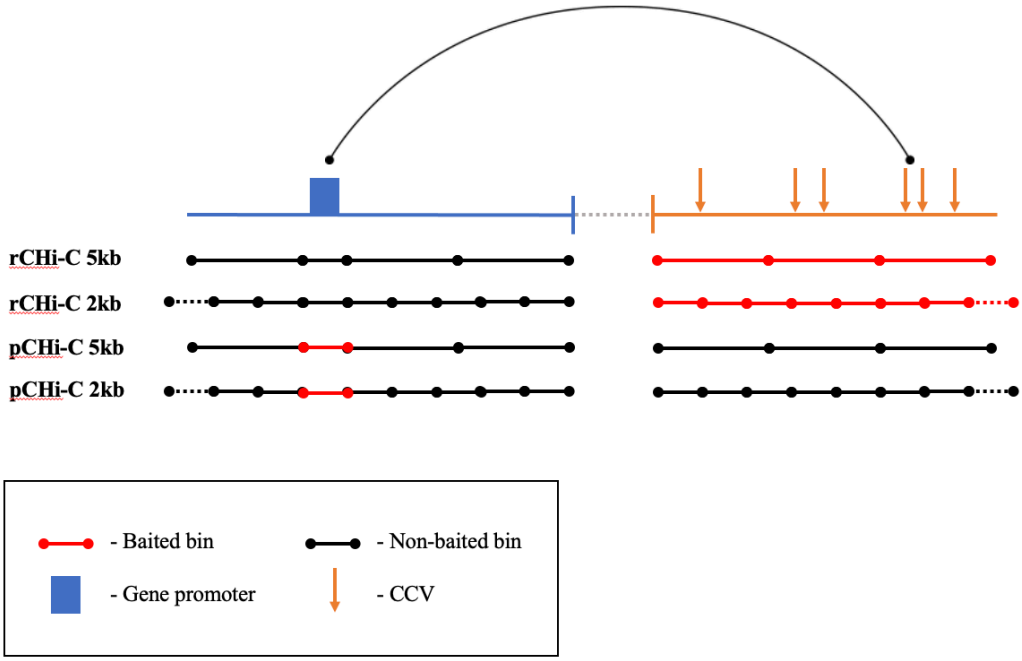

Figure 6.1: Distribution of bins in CHi-C libraries generated using the Dovetail Genomics Omni-C protocol.

Looking into the subsets of direct interaction peaks, more direct interaction peaks as well as 'direct' gene- and CCV-containing bins were observed in all the 2kb rCHi-C libraries, except the T-47D Arima library. One possible explanation is that this library had the lowest absolute number (and proportion) of on-target read pairs (Table 3.1) and, therefore, there was not enough power for the 2 kb analysis.

In conclusion, I observed that the choice of analysis has a major effect on the data. In the absence of a ground truth, i.e., a data set in which the true target genes and causal variants are known, it is difficult to know which is the 'best' analysis, but careful considerations are required when making a choice. Future studies investigating various options in-depth would be beneficial.

6.2. Limitations

In this study, I mainly focused on direct interaction peaks, i.e., interaction peaks where a bin colocalising with a gene promoter formed a direct interaction peak with a CCV-containing bin. However, it is important to remember that simply finding an interaction between a CCV and gene promoter does not on its own infer causality. It is likely that only a subset of the interacting variants has an effect on transcription factor binding or enhancer activity, some may influence transcription via a different mechanism, most will have no effect at all (bystanders). CHi-C can only identify interacting regions, but it cannot provide any information about the functional nature of the interaction. As a result, functional follow up studies (e.g., luciferase reporter assays or targeted CRISPR approaches) are required to investigate potential regulatory effects of these interactions.

It is possible that some interactions between the CCVs and putative target genes were missed due to the array coverage issues (lack of suitable baits). False negatives may occur due to short-range contact constraints or the transient nature of regulatory chromatin interactions. In addition, regulatory interactions appear to be cell type specific. In this study, I focused on luminal epithelial cells and fibroblasts, however, risk association at some signals may be mediated by other cell types (and/or stroma components), such as myoepithelial cells, adipocytes or immune cells. False positives can result from crosslinking artefacts, and it is possible that some interactions may be cell culture condition dependent.

The selection of CCVs used in this study is another important consideration. For the analysis, I used a list of 5,117 CCVs published in the latest BCAC paper[67], where they defined CCVs as all variants with a *p* value for association within two orders of magnitude of the index SNP. Although this approach has been consistently used in the breast cancer GWAS, it is not a convention found in other GWAS fields, and there is no real justification for this arbitrary cut-off.

This approach might work well for signals where there is a clear, individual variant associated with risk, such as signal 1 at the 2q35 locus, where rs4442975 is the index and the only SNP at the signal, and for which subsequent studies have confirmed rs4442975 as the functional variant[101, 114]. However, at some signals comprising many correlated variants, risk might be mediated by multiple variants of modest effect that are correlated with one another. In such cases, rather than a single strongly associated index variant

being the functional one, a haplotype of correlated variants may influence risk. In a recent study that used massively parallel reporter assays (MPRA) to systematically characterise the functional variants that drive eQTL and GWAS associations, Abell and colleagues found that at 17.7% of the eQTLs it was a haplotype of at least two variants in tight LD that was driving the associations[178]. In the extreme, they found two haplotypes comprising 13 correlated variants driving eQTL associations.

In addition, I designed my region capture array before the list of CCVs reported by the BCAC[67] became available. Therefore, my array was originally based on proxies, i.e., variants that were correlated with the index SNPs ($r^2 \geq 0.6$). As a result, the array regions did not capture all of the CCVs that mapped to the targeted 183 'strong-evidence' signals. Overall, my array included 3,383 out of 4,068 BCAC CCVs (83%), while the remaining 685 CCVs mapped outside of the array regions. For the excluded CCVs, I could not pick up interaction peaks in my rCHi-C data, and, therefore, cannot comment on their involvement in breast cancer risk mediation.

As a result of my work, I have now annotated BCAC fine-mapping regions with my CHi-C selected putative target genes (Table 5.5) and generated a version of the BCAC CCV data annotating which CCVs were involved in direct interaction peaks in primary cells. This thesis only reports numbers of CCVs that were involved in direct interaction peaks; it was not possible to provide a table of annotated variants due to the size limitations of the thesis. Overall, throughout the thesis, I focused on genes more than on CCVs for two main reasons: (i) the number of prioritised CCVs is much larger than the number of prioritised genes; (ii) there is much more published data available on individual genes than on individual CCVs. Realistically, within the scope of this project, all I could do with CCVs is to align them with markers of open chromatin, active histone modifications and TF-binding sites (and this is what I did for a few individual examples), but data in the relevant cell types (in primary cells and in breast fibroblasts) is quite limited.

6.3. Implications

There are two main ways in which CHi-C data can enhance our understanding of breast cancer biology. First, it allows prioritisation of a list of putative functional variants and target genes that warrant in-depth functional investigation. Follow up functional studies involving these genes and CCVs can help us to understand the mechanisms that increase breast cancer risk. Understanding these mechanisms might – in the longer term – inform

new ways of reducing breast cancer risk. In addition, cancer risk and progression may not be completely unrelated, so some of the identified genes could represent therapeutic targets. However, it is important to note that such theoretical developments are a long way from this current work. Secondly, identification of variants that affect breast cancer predisposition, can be used as part of polygenic risk score (PRS) models. PRS models can be used to predict likelihood of a polygenic trait by assigning weights to proxy measures of risk[179]. Fitting actual functional variants into PRS models will make them more intuitive, but not necessarily any more accurate, because these variants may not capture all of the risk at a certain locus quite as well as an index variant, and so should be tested against the original.

In summary, a high-throughput CHi-C analysis might contribute to on-going efforts to functionally characterise GWAS risk loci. Putative target genes and CCVs identified by CHi-C that are supported by additional data sources represent strong candidates for in-depth functional follow up studies. Therefore, my study represents an important resource for the breast cancer research community that can facilitate risk prediction, functional experimentation, and insights into breast cancer biology.

Bibliography

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209-49.
2. Eble JN, Tavassoli FA, Devilee P. Pathology and genetics of tumours of the breast and female genital organs: Iarc; 2003.
3. Waks AG, Winer EP. Breast Cancer Treatment: A Review. JAMA. 2019;321(3):288-300.
4. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, et al. Molecular portraits of human breast tumours. Nature. 2000;406(6797):747-52.
5. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, Rijn Mvd, Jeffrey SS, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proceedings of the National Academy of Sciences. 2001;98(19):10869-74.
6. Prat A, Perou CM. Deconstructing the molecular portraits of breast cancer. Mol Oncol. 2011;5(1):5-23.
7. Network TCGA. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490(7418):61-70.
8. Herschkowitz JI, Simin K, Weigman VJ, Mikaelian I, Usary J, Hu Z, Rasmussen KE, Jones LP, Assefnia S, Chandrasekharan S, et al. Identification of conserved gene expression features between murine mammary carcinoma models and human breast tumors. Genome Biol. 2007;8(5):R76.
9. Pommier RM, Sanlaville A, Tonon L, Kielbassa J, Thomas E, Ferrari A, Sertier A-S, Hollande F, Martinez P, Tissier A, et al. Comprehensive characterization of claudin-low breast tumors reflects the impact of the cell-of-origin on cancer evolution. Nature communications. 2020;11(1):3431.
10. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-7.
11. Chen X, Wang Q, Zhang Y, Xie Q, Tan X. Physical Activity and Risk of Breast Cancer: A Meta-Analysis of 38 Cohort Studies in 45 Study Reports. Value Health. 2019;22(1):104-28.
12. Ennour-Idrissi K, Maunsell E, Diorio C. Effect of physical activity on sex hormones in women: a systematic review and meta-analysis of randomized controlled trials. Breast cancer research : BCR. 2015;17(1):139.
13. Nieman DC, Wentz LM. The compelling link between physical activity and the body's defense system. J Sport Health Sci. 2019;8(3):201-17.
14. de Alcantara Borba D, da Silva Alves E, Rosa JPP, Facundo LA, Costa CMA, Silva AC, Narciso FV, Silva A, de Mello MT. Can IGF-1 Serum Levels Really be Changed by Acute Physical Exercise? A Systematic Review and Meta-Analysis. J Phys Act Health. 2020;17(5):575-84.
15. Kolb R, Zhang W. Obesity and Breast Cancer: A Case of Inflamed Adipose Tissue. Cancers (Basel). 2020;12(6).
16. James FR, Wootton S, Jackson A, Wiseman M, Copson ER, Cutress RI. Obesity in breast cancer--what is the risk factor? Eur J Cancer. 2015;51(6):705-20.
17. Zeinomar N, Knight JA, Genkinger JM, Phillips KA, Daly MB, Milne RL, Dite GS, Kehm RD, Liao Y, Southey MC, et al. Alcohol consumption, cigarette smoking, and

familial breast cancer risk: findings from the Prospective Family Study Cohort (ProF-SC). Breast cancer research : BCR. 2019;21(1):128.
18. Erol A, Ho AM, Winham SJ, Karpyak VM. Sex hormones in alcohol consumption: a systematic review of evidence. Addict Biol. 2019;24(2):157-69.
19. Coronado GD, Beasley J, Livaudais J. Alcohol consumption and the risk of breast cancer. Salud Publica Mex. 2011;53(5):440-7.
20. Terry PD, Rohan TE. Cigarette smoking and the risk of breast cancer in women: a review of the literature. Cancer Epidemiol Biomarkers Prev. 2002;11(10 Pt 1):953-71.
21. Poorolajal J, Heidarimoghis F, Karami M, Cheraghi Z, Gohari-Ensaf F, Shahbazi F, Zareie B, Ameri P, Sahraee F. Factors for the Primary Prevention of Breast Cancer: A Meta-Analysis of Prospective Cohort Studies. J Res Health Sci. 2021;21(3):e00520.
22. Atoum M, Alzoughool F. Vitamin D and Breast Cancer: Latest Evidence and Future Steps. Breast Cancer (Auckl). 2017;11:1178223417749816.
23. Huss L, Butt ST, Borgquist S, Elebro K, Sandsveden M, Rosendahl A, Manjer J. Vitamin D receptor expression in invasive breast tumors and breast cancer survival. Breast Cancer Research. 2019;21(1):84.
24. Fiolet T, Srour B, Sellem L, Kesse-Guyot E, Allès B, Méjean C, Deschasaux M, Fassier P, Latino-Martel P, Beslay M, et al. Consumption of ultra-processed foods and cancer risk: results from NutriNet-Santé prospective cohort. BMJ. 2018;360:k322.
25. Hill DA, Prossnitz ER, Royce M, Nibbe A. Temporal trends in breast cancer survival by race and ethnicity: A population-based cohort study. PloS one. 2019;14(10):e0224064.
26. Kurian AW, Fish K, Shema SJ, Clarke CA. Lifetime risks of specific breast cancer subtypes among women in four racial/ethnic groups. Breast cancer research : BCR. 2010;12(6):R99.
27. Benz CC. Impact of aging on the biology of breast cancer. Crit Rev Oncol/Hematol. 2008;66(1):65-74.
28. McGuire A, Brown JA, Malone C, McLaughlin R, Kerin MJ. Effects of age on the detection and management of breast cancer. Cancers (Basel). 2015;7(2):908-29.
29. Cancer CGoHFiB. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58 209 women with breast cancer and 101 986 women without the disease. The Lancet. 2001;358(9291):1389-99.
30. Albrektsen G, Heuch I, Hansen S, Kvåle G. Breast cancer risk by age at birth, time since birth and time intervals between births: exploring interaction effects. Br J Cancer. 2005;92(1):167-75.
31. Husby A, Wohlfahrt J, Øyen N, Melbye M. Pregnancy duration and breast cancer risk. Nature communications. 2018;9(1):4255.
32. Callihan EB, Gao D, Jindal S, Lyons TR, Manthey E, Edgerton S, Urquhart A, Schedin P, Borges VF. Postpartum diagnosis demonstrates a high risk for metastasis and merits an expanded definition of pregnancy-associated breast cancer. Breast Cancer Research and Treatment. 2013;138(2):549-59.
33. Ursin G, Bernstein L, Lord SJ, Karim R, Deapen D, Press MF, Daling JR, Norman SA, Liff JM, Marchbanks PA, et al. Reproductive factors and subtypes of breast cancer defined by hormone receptor and histology. Br J Cancer. 2005;93(3):364-71.
34. Cancer CGoHFiB. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50 302 women with breast cancer and 96 973 women without the disease. The Lancet. 2002;360(9328):187-95.
35. Cancer CGoHFiB. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. Lancet Oncol. 2012;13(11):1141-51.

36. Narod SA. Hormone replacement therapy and the risk of breast cancer. Nat Rev Clin Oncol. 2011;8(11):669-76.
37. Boyd NF, Rommens JM, Vogt K, Lee V, Hopper JL, Yaffe MJ, Paterson AD. Mammographic breast density as an intermediate phenotype for breast cancer. Lancet Oncol. 2005;6(10):798-808.
38. Boyd N, Berman H, Zhu J, Martin LJ, Yaffe MJ, Chavez S, Stanisz G, Hislop G, Chiarelli AM, Minkin S, et al. The origins of breast cancer associated with mammographic density: a testable biological hypothesis. Breast cancer research : BCR. 2018;20(1):17.
39. Boyd NF, Guo H, Martin LJ, Sun L, Stone J, Fishell E, Jong RA, Hislop G, Chiarelli A, Minkin S, et al. Mammographic density and the risk and detection of breast cancer. N Engl J Med. 2007;356(3):227-36.
40. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. Cancer Epidemiol Biomarkers Prev. 2006;15(6):1159-69.
41. Pettersson A, Graff RE, Ursin G, Santos Silva ID, McCormack V, Baglietto L, Vachon C, Bakker MF, Giles GG, Chia KS, et al. Mammographic density phenotypes and risk of breast cancer: a meta-analysis. J Natl Cancer Inst. 2014;106(5).
42. Pettersson A, Hankinson SE, Willett WC, Lagiou P, Trichopoulos D, Tamimi RM. Nondense mammographic area and risk of breast cancer. Breast Cancer Research. 2011;13(5):R100.
43. Hanahan D. Hallmarks of Cancer: New Dimensions. Cancer Discovery. 2022;12(1):31-46.
44. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC. Linkage of early-onset familial breast cancer to chromosome 17q21. Science (New York, NY). 1990;250(4988):1684-9.
45. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, Nguyen K, Seal S, Tran T, Averill D, et al. Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. Science (New York, NY). 1994;265(5181):2088-90.
46. McClain MR, Palomaki GE, Nathanson KL, Haddow JE. Adjusting the estimated proportion of breast cancer cases associated with BRCA1 and BRCA2 mutations: public health implications. Genet Med. 2005;7(1):28-33.
47. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. Br J Cancer. 2000;83(10):1301-8.
48. Kechagioglou P, Papi RM, Provatopoulou X, Kalogera E, Papadimitriou E, Grigoropoulos P, Nonni A, Zografos G, Kyriakidis DA, Gounaris A. Tumor suppressor PTEN in breast cancer: heterozygosity, mutations and protein expression. Anticancer Res. 2014;34(3):1387-400.
49. Shahbandi A, Nguyen HD, Jackson JG. TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines. Trends Cancer. 2020;6(2):98-110.
50. Corso G, Intra M, Trentin C, Veronesi P, Galimberti V. CDH1 germline mutations and hereditary lobular breast cancer. Fam Cancer. 2016;15(2):215-9.
51. Chen J, Lindblom A. Germline mutation screening of the STK11/LKB1 gene in familial breast cancer with LOH on 19p. Clin Genet. 2000;57(5):394-7.
52. Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, et al. Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat Genet. 2002;31(1):55-9.
53. Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, Spanova K, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nat Genet. 2006;38(8):873-5.

54. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilake H, Ahmed M, Spanova K, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat Genet. 2006;38(11):1239-41.
55. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid S, Spanova K, Barfoot R, Chagtai T, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. Nat Genet. 2007;39(2):165-7.
56. Consortium IHGS. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.
57. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, Vollger MR, Altemose N, Uralsky L, Gershman A, et al. The complete sequence of a human genome. Science (New York, NY). 2022;376(6588):44-53.
58. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.
59. Børsting C, Morling N. Single-Nucleotide Polymorphisms. In: Siegel JA, Saukko PJ, Houck MM, editors. Encyclopedia of Forensic Sciences (Second Edition). Waltham: Academic Press; 2013. p. 233-8.
60. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015;526(7571):75-81.
61. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, Yang J. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5-22.
62. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. Complement factor H polymorphism in age-related macular degeneration. Science (New York, NY). 2005;308(5720):385-9.
63. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896-d901.
64. Sakoda LC, Jorgenson E, Witte JS. Turning of COGS moves forward findings for hormonally mediated cancers. Nat Genet. 2013;45(4):345-8.
65. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, Casey G, Hunter DJ, Sellers TA, Gruber SB, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev. 2017;26(1):126-35.
66. Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551(7678):92-4.
67. Fachal L, Aschard H, Beesley J, Barnes DR, Allen J, Kar S, Pooley KA, Dennis J, Michailidou K, Turman C, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet. 2020;52(1):56-73.
68. Long J, Cai Q, Sung H, Shi J, Zhang B, Choi JY, Wen W, Delahanty RJ, Lu W, Gao YT, et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS Genet. 2012;8(2):e1002532.
69. Chen F, Chen GK, Stram DO, Millikan RC, Ambrosone CB, John EM, Bernstein L, Zheng W, Palmer JR, Hu JJ, et al. A genome-wide association study of breast cancer in women of African ancestry. Hum Genet. 2013;132(1):39-48.
70. Huo D, Feng Y, Haddad S, Zheng Y, Yao S, Han Y-J, Ogundiran TO, Adebamowo C, Ojengbede O, Falusi AG, et al. Genome-wide association studies in

women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. Hum Mol Genet. 2016;25(21):4835-46.
71. Zheng W, Zhang B, Cai Q, Sung H, Michailidou K, Shi J, Choi JY, Long J, Dennis J, Humphreys MK, et al. Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. Hum Mol Genet. 2013;22(12):2539-50.
72. Cai Q, Zhang B, Sung H, Low SK, Kweon SS, Lu W, Shi J, Long J, Wen W, Choi JY, et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. Nat Genet. 2014;46(8):886-90.
73. Chen H, Fan S, Stone J, Thompson DJ, Douglas J, Li S, Scott C, Bolla MK, Wang Q, Dennis J, et al. Genome-wide and transcriptome-wide association studies of mammographic density phenotypes reveal novel loci. Breast Cancer Research. 2022;24(1):27.
74. Lindstrom S, Thompson DJ, Paterson AD, Li J, Gierach GL, Scott C, Stone J, Douglas JA, dos-Santos-Silva I, Fernandez-Navarro P, et al. Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk. Nature communications. 2014;5:5303.
75. Sieh W, Rothstein JH, Klein RJ, Alexeeff SE, Sakoda LC, Jorgenson E, McBride RB, Graff RE, McGuire V, Achacoso N, et al. Identification of 31 loci for mammographic density phenotypes and their associations with breast cancer risk. Nature communications. 2020;11(1):5116.
76. He C, Kraft P, Chen C, Buring JE, Paré G, Hankinson SE, Chanock SJ, Ridker PM, Hunter DJ, Chasman DI. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. Nat Genet. 2009;41(6):724-8.
77. Ahsan H, Halpern J, Kibriya MG, Pierce BL, Tong L, Gamazon E, McGuire V, Felberg A, Shi J, Jasmine F, et al. A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. Cancer Epidemiol Biomarkers Prev. 2014;23(4):658-69.
78. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet. 2011;43(6):513-8.
79. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(23):9362-7.
80. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, Reynolds AP, Sandstrom R, Qu H, Brody J, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science (New York, NY). 2012;337(6099):1190-5.
81. Dekker J, Rippe K, Dekker M, Kleckner N. Capturing chromosome conformation. Science (New York, NY). 2002;295(5558):1306-11.
82. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet. 2006;38(11):1348-54.
83. Zhao Z, Tavoosidana G, Sjolinder M, Gondor A, Mariano P, Wang S, Kanduri C, Lezcano M, Sandhu KS, Singh U, et al. Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat Genet. 2006;38(11):1341-7.
84. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, et al. Chromosome Conformation Capture Carbon Copy

(5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 2006;16(10):1299-309.
85. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science (New York, NY). 2009;326(5950):289-93.
86. Kaplan N, Dekker J. High-throughput genome scaffolding from in vivo DNA interaction frequency. Nat Biotechnol. 2013;31(12):1143-7.
87. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, et al. A chromosome conformation capture ordered sequence of the barley genome. Nature. 2017;544(7651):427-33.
88. Harewood L, Kishore K, Eldridge MD, Wingett S, Pearson D, Schoenfelder S, Collins VP, Fraser P. Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours. Genome Biology. 2017;18(1):125.
89. Belaghzal H, Dekker J, Gibcus JH. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods. 2017;123:56-65.
90. Gavrilov A, Razin SV, Cavalli G. In vivo formaldehyde cross-linking: it is time for black box analysis. Brief Funct Genomics. 2015;14(2):163-5.
91. Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: a comprehensive technique to capture the conformation of genomes. Methods. 2012;58(3):268-76.
92. van Berkum NL, Lieberman-Aiden E, Williams L, Imakaev M, Gnirke A, Mirny LA, Dekker J, Lander ES. Hi-C: a method to study the three-dimensional architecture of genomes. J Vis Exp. 2010(39).
93. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665-80.
94. Ma W, Ay F, Lee C, Gulsoy G, Deng X, Cook S, Hesson J, Cavanaugh C, Ware CB, Krumm A, et al. Using DNase Hi-C techniques to map global and local three-dimensional genome architecture at high resolution. Methods. 2018;142:59-73.
95. Ramani V, Cusanovich DA, Hause RJ, Ma W, Qiu R, Deng X, Blau CA, Disteche CM, Noble WS, Shendure J, et al. Mapping 3D genome architecture through in situ DNase Hi-C. Nat Protoc. 2016;11(11):2104-21.
96. Hsieh TH, Weiner A, Lajoie B, Dekker J, Friedman N, Rando OJ. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. Cell. 2015;162(1):108-19.
97. Nagano T, Várnai C, Schoenfelder S, Javierre BM, Wingett SW, Fraser P. Comparison of Hi-C results using in-solution versus in-nucleus ligation. Genome Biol. 2015;16(1):175.
98. Göndör A, Rougier C, Ohlsson R. High-resolution circular chromosome conformation capture assay. Nature Protocols. 2008;3(2):303-13.
99. Naumova N, Smith EM, Zhan Y, Dekker J. Analysis of long-range chromatin interactions using Chromosome Conformation Capture. Methods. 2012;58(3):192-203.
100. Lajoie BR, Dekker J, Kaplan N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. Methods. 2015;72:65-75.
101. Dryden NH, Broome LR, Dudbridge F, Johnson N, Orr N, Schoenfelder S, Nagano T, Andrews S, Wingett S, Kozarewa I, et al. Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. Genome Res. 2014;24(11):1854-68.

102. Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. Target-enrichment strategies for next-generation sequencing. Nat Methods. 2010;7(2):111-8.
103. Orlando G, Kinnersley B, Houlston RS. Capture Hi-C Library Generation and Analysis to Detect Chromatin Interactions. Curr Protoc Hum Genet. 2018:e63.
104. Akgol Oksuz B, Yang L, Abraham S, Venev SV, Krietenstein N, Parsi KM, Ozadam H, Oomen ME, Nand A, Mao H, et al. Systematic evaluation of chromosome conformation capture assays. Nat Methods. 2021;18(9):1046-55.
105. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med. 2009;15(8):907-13.
106. Valkenburg KC, de Groot AE, Pienta KJ. Targeting the tumour stroma to improve cancer therapy. Nat Rev Clin Oncol. 2018;15(6):366-81.
107. Pickrell JK, Berisa T, Liu JZ, Segurel L, Tung JY, Hinds DA. Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet. 2016;48(7):709-17.
108. Holgersen EM, Gillespie A, Leavy OC, Baxter JS, Zvereva A, Muirhead G, Johnson N, Sipos O, Dryden NH, Broome LR, et al. Identifying high-confidence capture Hi-C interactions using CHiCANE. Nature Protocols. 2021;16(4):2257-85.
109. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
110. Vavouri T, McEwen GK, Woolfe A, Gilks WR, Elgar G. Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. Trends Genet. 2006;22(1):5-10.
111. Sanyal A, Lajoie BR, Jain G, Dekker J. The long-range interaction landscape of gene promoters. Nature. 2012;489(7414):109-13.
112. Baxter JS, Johnson N, Tomczyk K, Gillespie A, Maguire S, Brough R, Fachal L, Michailidou K, Bolla MK, Wang Q, et al. Functional annotation of the 2q35 breast cancer risk locus implicates a structural variant in influencing activity of a long-range enhancer element. Am J Hum Genet. 2021;108(7):1190-203.
113. Wyszynski A, Hong CC, Lam K, Michailidou K, Lytle C, Yao S, Zhang Y, Bolla MK, Wang Q, Dennis J, et al. An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. Hum Mol Genet. 2016;25(17):3863-76.
114. Ghoussaini M, Edwards SL, Michailidou K, Nord S, Cowper-Sal Lari R, Desai K, Kar S, Hillman KM, Kaufmann S, Glubb DM, et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nature communications. 2014;4:4999.
115. de Laat W, Duboule D. Topology of mammalian developmental enhancers and their regulatory landscapes. Nature. 2013;502(7472):499-506.
116. Bonev B, Cavalli G. Organization and function of the 3D genome. Nature Reviews Genetics. 2016;17(11):661-78.
117. Beesley J, Sivakumaran H, Moradi Marjaneh M, Lima LG, Hillman KM, Kaufmann S, Tuano N, Hussein N, Ham S, Mukhopadhyay P, et al. Chromatin interactome mapping at 139 independent breast cancer risk signals. Genome Biology. 2020;21(1):8.
118. Choy M-K, Javierre BM, Williams SG, Baross SL, Liu Y, Wingett SW, Akbarov A, Wallace C, Freire-Pritchett P, Rugg-Gunn PJ, et al. Promoter interactome of human embryonic stem cell-derived cardiomyocytes connects GWAS regions to cardiac gene networks. Nature communications. 2018;9(1):2526.

119. Montefiori LE, Sobreira DR, Sakabe NJ, Aneas I, Joslin AC, Hansen GT, Bozek G, Moskowitz IP, McNally EM, Nóbrega MA. A promoter interaction map for cardiovascular disease genetics. eLife. 2018;7:e35788.
120. Selvarajan I, Toropainen A, Garske KM, López Rodríguez M, Ko A, Miao Z, Kaminska D, Õunap K, Örd T, Ravindran A, et al. Integrative analysis of liver-specific non-coding regulatory SNPs associated with the risk of coronary artery disease. Am J Hum Genet. 2021;108(3):411-30.
121. Chandra V, Bhattacharyya S, Schmiedel BJ, Madrigal A, Gonzalez-Colin C, Fotsing S, Crinklaw A, Seumois G, Mohammadi P, Kronenberg M, et al. Promoter-interacting expression quantitative trait loci are enriched for functional genetic variants. Nat Genet. 2021;53(1):110-9.
122. Yan L, Lin M, Pan S, Assaraf YG, Wang Z-w, Zhu X. Emerging roles of F-box proteins in cancer drug resistance. Drug Resistance Updates. 2020;49:100673.
123. Zhou H, Liu Y, Zhu R, Ding F, Wan Y, Li Y, Liu Z. FBXO32 suppresses breast cancer tumorigenesis through targeting KLF4 to proteasomal degradation. Oncogene. 2017;36(23):3312-21.
124. Orr N, Dudbridge F, Dryden N, Maguire S, Novo D, Perrakis E, Johnson N, Ghoussaini M, Hopper JL, Southey MC, et al. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Hum Mol Genet. 2015;24(10):2966-84.
125. Schwab M. Amplification of oncogenes in human cancer cells. Bioessays. 1998;20(6):473-9.
126. Velasco-Velázquez MA, Li Z, Casimiro M, Loro E, Homsi N, Pestell RG. Examining the role of cyclin D1 in breast cancer. Future Oncol. 2011;7(6):753-65.
127. Fusté NP, Fernández-Hernández R, Cemeli T, Mirantes C, Pedraza N, Rafel M, Torres-Rosell J, Colomina N, Ferrezuelo F, Dolcet X, et al. Cytoplasmic cyclin D1 regulates cell invasion and metastasis through the phosphorylation of paxillin. Nature communications. 2016;7:11581.
128. Janssen JW, Cuny M, Orsetti B, Rodriguez C, Vallés H, Bartram CR, Schuuring E, Theillet C. MYEOV: a candidate gene for DNA amplification events occurring centromeric to CCND1 in breast cancer. Int J Cancer. 2002;102(6):608-14.
129. Fang L, Wu S, Zhu X, Cai J, Wu J, He Z, Liu L, Zeng M, Song E, Li J, et al. MYEOV functions as an amplified competing endogenous RNA in promoting metastasis by activating TGF-β pathway in NSCLC. Oncogene. 2019;38(6):896-912.
130. Romualdo Cardoso S, Gillespie A, Haider S, Fletcher O. Functional annotation of breast cancer risk loci: current progress and future directions. Br J Cancer. 2022;126(7):981-93.
131. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2011;40(D1):D930-D4.
132. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, Karczewski KJ, Park J, Hitz BC, Weng S, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22(9):1790-7.
133. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
134. Melnikov A, Murugan A, Zhang X, Tesileanu T, Wang L, Rogov P, Feizi S, Gnirke A, Callan CG, Jr., Kinney JB, et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol. 2012;30(3):271-7.
135. Arnold CD, Gerlach D, Stelzer C, Boryń Ł M, Rath M, Stark A. Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science (New York, NY). 2013;339(6123):1074-7.

136. Huang X, Liu X, Du B, Liu X, Xue M, Yan Q, Wang X, Wang Q. LncRNA LINC01305 promotes cervical cancer progression through KHSRP and exosome-mediated transfer. Aging (Albany NY). 2021;13(15):19230-42.
137. Xu Y, Zhu H, Ma H, Yuan L, Hu Q, Yang L. LINC01305 inhibits malignant progression of cervical cancer via miR-129-5p/Sox4 axis. Am J Transl Res. 2020;12(11):7581-92.
138. Zhang JW, Rubio V, Zheng S, Shi ZZ. Knockdown of OLA1, a regulator of oxidative stress response, inhibits motility and invasion of breast cancer cells. J Zhejiang Univ Sci B. 2009;10(11):796-804.
139. Liu J, Miao X, Xiao B, Huang J, Tao X, Zhang J, Zhao H, Pan Y, Wang H, Gao G, et al. Obg-Like ATPase 1 Enhances Chemoresistance of Breast Cancer via Activation of TGF-β/Smad Axis Cascades. Front Pharmacol. 2020;11:666.
140. Min A, Jang H, Kim S, Lee KH, Kim DK, Suh KJ, Yang Y, Elvin P, O'Connor MJ, Im SA. Androgen Receptor Inhibitor Enhances the Antitumor Effect of PARP Inhibitor in Breast Cancer Cells by Modulating DNA Damage Response. Mol Cancer Ther. 2018;17(12):2507-18.
141. Essegian D, Khurana R, Stathias V, Schürer SC. The Clinical Kinase Index: A Method to Prioritize Understudied Kinases as Drug Targets for the Treatment of Cancer. Cell Rep Med. 2020;1(7):100128.
142. Miwa T, Kanda M, Shimizu D, Umeda S, Sawaki K, Tanaka H, Tanaka C, Hattori N, Hayashi M, Yamada S, et al. Hepatic metastasis of gastric cancer is associated with enhanced expression of ethanolamine kinase 2 via the p53-Bcl-2 intrinsic apoptosis pathway. Br J Cancer. 2021;124(8):1449-60.
143. Mehta GA, Khanna P, Gatza ML. Emerging Role of SOX Proteins in Breast Cancer Development and Maintenance. Journal of mammary gland biology and neoplasia. 2019;24(3):213-30.
144. Jin X, Shao X, Pang W, Wang Z, Huang J. Sex-determining Region Y-box transcription factor 13 promotes breast cancer cell proliferation and glycolysis by activating the tripartite motif containing 11-mediated Wnt/β-catenin signaling pathway. Bioengineered. 2022;13(5):13033-44.
145. Veltmaat JM, Relaix F, Le LT, Kratochwil K, Sala FG, van Veelen W, Rice R, Spencer-Dene B, Mailleux AA, Rice DP, et al. Gli3-mediated somitic Fgf10 expression gradients are required for the induction and patterning of mammary epithelium along the embryonic axes. Development. 2006;133(12):2325-35.
146. Mailleux AA, Spencer-Dene B, Dillon C, Ndiaye D, Savona-Baron C, Itoh N, Kato S, Dickson C, Thiery JP, Bellusci S. Role of FGF10/FGFR2b signaling during mammary gland development in the mouse embryo. Development. 2002;129(1):53-60.
147. Theodorou V, Boer M, Weigelt B, Jonkers J, van der Valk M, Hilkens J. Fgf10 is an oncogene activated by MMTV insertional mutagenesis in mouse mammary tumors and overexpressed in a subset of human breast carcinomas. Oncogene. 2004;23(36):6047-55.
148. Grigoriadis A, Mackay A, Reis-Filho JS, Steele D, Iseli C, Stevenson BJ, Jongeneel CV, Valgeirsson H, Fenwick K, Iravani M, et al. Establishment of the epithelial-specific transcriptome of normal and malignant human breast cells based on MPSS and array expression data. Breast cancer research : BCR. 2006;8(5):R56.
149. Abolhassani A, Riazi GH, Azizi E, Amanpour S, Muhammadnejad S, Haddadi M, Zekri A, Shirkoohi R. FGF10: Type III Epithelial Mesenchymal Transition and Invasion in Breast Cancer Cell Lines. J Cancer. 2014;5(7):537-47.
150. Ghoussaini M, French JD, Michailidou K, Nord S, Beesley J, Canisus S, Hillman KM, Kaufmann S, Sivakumaran H, Moradi Marjaneh M, et al. Evidence that the 5p12 Variant rs10941679 Confers Susceptibility to Estrogen-Receptor-Positive Breast Cancer through FGF10 and MRPS30 Regulation. Am J Hum Genet. 2016;99(4):903-11.

151. Xiang X, Deng Z, Zhuang X, Ju S, Mu J, Jiang H, Zhang L, Yan J, Miller D, Zhang HG. Grhl2 determines the epithelial phenotype of breast cancers and promotes tumor progression. PloS one. 2012;7(12):e50781.
152. Wang Y, Xu H, Zhu B, Qiu Z, Lin Z. Systematic identification of the key candidate genes in breast cancer stroma. Cell Mol Biol Lett. 2018;23:44.
153. Feng L-y, Li L. Low expression of NCALD is associated with chemotherapy resistance and poor prognosis in epithelial ovarian cancer. Journal of Ovarian Research. 2020;13(1):35.
154. Pan S, Deng Y, Fu J, Zhang Y, Zhang Z, Qin X. N6-methyladenosine upregulates miR-181d-5p in exosomes derived from cancer-associated fibroblasts to inhibit 5-FU sensitivity by targeting NCALD in colorectal cancer. Int J Oncol. 2022;60(2):14.
155. Wang H, Wei H, Wang J, Li L, Chen A, Li Z. MicroRNA-181d-5p-Containing Exosomes Derived from CAFs Promote EMT by Regulating CDX2/HOXA5 in Breast Cancer. Mol Ther Nucleic Acids. 2020;19:654-67.
156. Wang MC, Li CL, Cui J, Jiao M, Wu T, Jing LI, Nan KJ. BMI-1, a promising therapeutic target for human cancer. Oncol Lett. 2015;10(2):583-8.
157. De Luca A, Sanna F, Sallese M, Ruggiero C, Grossi M, Sacchetta P, Rossi C, De Laurenzi V, Di Ilio C, Favaloro B. Methionine sulfoxide reductase A down-regulation in human breast cancer cells results in a more aggressive phenotype. Proceedings of the National Academy of Sciences of the United States of America. 2010;107(43):18628-33.
158. Morel A-P, Ginestier C, Pommier RM, Cabaud O, Ruiz E, Wicinski J, Devouassoux-Shisheboran M, Combaret V, Finetti P, Chassot C, et al. A stemness-related ZEB1–MSRB3 axis governs cellular pliancy and breast cancer genome stability. Nat Med. 2017;23(5):568-78.
159. Lee SH, Lee S, Du J, Jain K, Ding M, Kadado AJ, Atteya G, Jaji Z, Tyagi T, Kim WH, et al. Mitochondrial MsrB2 serves as a switch and transducer for mitophagy. EMBO Mol Med. 2019;11(8):e10409.
160. Picot CR, Perichon M, Cintrat J-C, Friguet B, Petropoulos I. The peptide methionine sulfoxide reductases, MsrA and MsrB (hCBS-1), are downregulated during replicative senescence of human WI-38 fibroblasts. FEBS Lett. 2004;558(1-3):74-8.
161. Oien DB, Moskovitz J. Genetic regulation of longevity and age-associated diseases through the methionine sulfoxide reductase system. Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease. 2019;1865(7):1756-62.
162. Lu Y, Li J, Cheng D, Parameswaran B, Zhang S, Jiang Z, Yew PR, Peng J, Ye Q, Hu Y. The F-box protein FBXO44 mediates BRCA1 ubiquitination and degradation. J Biol Chem. 2012;287(49):41014-22.
163. Thompson LL, Rutherford KA, Lepage CC, McManus KJ. Aberrant SKP1 Expression: Diverse Mechanisms Impacting Genome and Chromosome Stability. Front Cell Dev Biol. 2022;10:859582.
164. Brooks DLP, Schwab LP, Krutilina R, Parke DN, Sethuraman A, Hoogewijs D, Schörg A, Gotwald L, Fan M, Wenger RH, et al. ITGA6 is directly regulated by hypoxia-inducible factors and enriches for cancer stem cell activity and invasion in metastatic breast cancer models. Molecular Cancer. 2016;15(1):26.
165. Hu T, Zhou R, Zhao Y, Wu G. Integrin α6/Akt/Erk signaling is essential for human breast cancer resistance to radiotherapy. Scientific reports. 2016;6(1):33376.
166. Kim BG, An HJ, Kang S, Choi YP, Gao MQ, Park H, Cho NH. Laminin-332-rich tumor microenvironment for tumor invasion in the interface zone of breast cancer. Am J Pathol. 2011;178(1):373-81.
167. Chen H, Qu J, Huang X, Kurundkar A, Zhu L, Yang N, Venado A, Ding Q, Liu G, Antony VB, et al. Mechanosensing by the α6-integrin confers an invasive fibroblast phenotype and mediates lung fibrosis. Nature communications. 2016;7:12564.

168. Yuan J, Li P, Pan H, Li Y, Xu Q, Xu T, Ji X, Liu Y, Yao W, Han L, et al. miR-542-5p Attenuates Fibroblast Activation by Targeting Integrin α6 in Silica-Induced Pulmonary Fibrosis. Int J Mol Sci. 2018;19(12).
169. Stevens TA, Meech R. BARX2 and estrogen receptor-α (ESR1) coordinately regulate the production of alternatively spliced ESR1 isoforms and control breast cancer cell growth and invasion. Oncogene. 2006;25(39):5426-35.
170. Cavalli G, Misteli T. Functional implications of genome topology. Nat Struct Mol Biol. 2013;20(3):290-9.
171. Pennacchio LA, Bickmore W, Dean A, Nobrega MA, Bejerano G. Enhancers: five essential questions. Nat Rev Genet. 2013;14(4):288-95.
172. Yang J, McGovern A, Martin P, Duffus K, Ge X, Zarrineh P, Morris AP, Adamson A, Fraser P, Rattray M, et al. Analysis of chromatin organization and gene expression in T cells identifies functional genes for rheumatoid arthritis. Nature communications. 2020;11(1):4402.
173. Mumbach MR, Rubin AJ, Flynn RA, Dai C, Khavari PA, Greenleaf WJ, Chang HY. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. Nat Methods. 2016;13(11):919-22.
174. Datlinger P, Rendeiro AF, Schmidl C, Krausgruber T, Traxler P, Klughammer J, Schuster LC, Kuchler A, Alpar D, Bock C. Pooled CRISPR screening with single-cell transcriptome readout. Nat Methods. 2017;14(3):297-301.
175. Chovanec P, Collier AJ, Krueger C, Várnai C, Semprich CI, Schoenfelder S, Corcoran AE, Rugg-Gunn PJ. Widespread reorganisation of pluripotent factor binding and gene regulatory interactions between human pluripotent states. Nature communications. 2021;12(1):2098.
176. Novo CL, Javierre BM, Cairns J, Segonds-Pichon A, Wingett SW, Freire-Pritchett P, Furlan-Magaril M, Schoenfelder S, Fraser P, Rugg-Gunn PJ. Long-Range Enhancer Interactions Are Prevalent in Mouse Embryonic Stem Cells and Are Reorganized upon Pluripotent State Transition. Cell Rep. 2018;22(10):2615-27.
177. Freire-Pritchett P, Schoenfelder S, Várnai C, Wingett SW, Cairns J, Collier AJ, García-Vílchez R, Furlan-Magaril M, Osborne CS, Fraser P, et al. Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells. Elife. 2017;6.
178. Abell NS, DeGorter MK, Gloudemans MJ, Greenwald E, Smith KS, He Z, Montgomery SB. Multiple causal variants underlie genetic associations in humans. Science (New York, NY). 2022;375(6586):1247-54.
179. Mavaddat N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst. 2015;107(5).

Appendix A

Regions targeted by the rCHi-C capture array.

Array design is described in Section 2.5. Chr – chromosome; Region start – start coordinate of the final array region; Region end – end coordinate of the final array region; SNP type – index SNP type (bc – breast cancer risk SNP, md – mammographic density SNP, size – breast size SNP, credvar – credible variant from Michailidou et al., 2017). All coordinates are in GRCh38/hg38. The focus of my thesis is the breast cancer risk regions; MD and breast size loci were also included on the array but will be the subject of a subsequent analysis.

Chr	Cytoband	Region start	Region end	Index SNP	SNP type
chr1	p36.22	10,489,676	10,492,573	rs657244	bc
chr1	p36.13	18,467,486	18,484,134	rs2992756	bc
chr1	p34.1, p33	46,209,587	46,418,560	rs12039667	bc
chr1	p22.3	87,682,256	87,731,574	rs7541276	bc
chr1	p22.2	87,948,543	87,976,953	rs11583393	bc
chr1	p13.2	113,624,200	113,636,095	rs11552449	credvar
chr1	p13.2	113,652,833	113,910,769	rs11102701	bc
chr1	p12	117,638,307	117,731,142	rs7529522	bc
chr1	p12	118,194,960	118,296,940	rs2359714	size
chr1	q21.1	143,901,312	143,986,674	rs201000337	bc
chr1	p11.2	121,537,792	121,541,057	rs11249433	bc
chr1	q21.1	145,720,339	145,849,216	rs143384623	bc
chr1	q21.1	145,624,176	145,710,306	rs200366104	bc
chr1	q21.2	149,905,982	150,028,456	rs12048493	bc
chr1	q22	155,165,982	155,232,616	rs1057941	bc
chr1	q22	155,286,651	155,310,810	rs4971059	credvar
chr1	q22	155,440,841	155,447,640	rs4971059	credvar
chr1	q22	155,580,577	155,589,158	rs4971059	credvar
chr1	q32.1	201,460,879	201,490,909	rs35383942	bc
chr1	q32.1	201,808,031	201,922,777	rs34091558	size
chr1	q32.1	202,202,758	202,239,749	rs4950774	bc
chr1	q32.1	203,796,228	203,982,201	rs59867004	bc
chr1	q32.1	204,478,666	204,622,747	rs4951401	bc
chr1	q41	217,003,003	217,049,973	rs11117754	bc
chr1	q43	241,854,596	241,874,165	rs72755295	bc
chr2	p25.3	598,965	658,786	rs62105303	size
chr2	p24.1	19,100,567	19,262,264	rs10184522	bc
chr2	p23.3	24,837,379	24,944,468	rs2384061	bc
chr2	p23.2	28,711,856	28,965,615	rs71403627	bc
chr2	q14.2	120,222,322	120,318,919	rs11448973	bc
chr2	q14.2	120,321,813	120,334,959	rs17625845	bc and size
chr2	q14.2	120,384,930	120,440,091	rs4076654	bc
chr2	q14.2	120,474,096	120,490,342	rs4849879	bc
chr2	q31.1	171,512,416	171,586,502	rs13020413	bc
chr2	q31.1	172,100,816	172,116,344	rs2016394	bc
chr2	q31.1	173,333,342	173,394,869	rs7589172	bc
chr2	q33.1	201,122,205	201,213,352	rs13015648	bc
chr2	q33.1	201,226,789	201,322,837	rs3769821	bc
chr2	q33.1	201,336,626	201,344,457	rs1830298	credvar
chr2	q35	217,011,844	217,061,244	rs4442975	bc
chr2	q35	217,084,532	217,115,592	rs138522813	bc
chr2	q35	217,385,406	217,494,724	rs5838651	bc
chr3	p26.1	4,685,845	4,721,977	rs6787391	bc
chr3	p24.1	27,225,136	27,355,050	rs36078735, rs1352944	bc, bc
chr3	p24.1	30,625,036	30,656,156	rs35263707	bc
chr3	p21.31	46,814,627	46,863,643	rs56387622	bc
chr3	p14.1	63,836,578	64,080,223	rs555060306	bc
chr3	p12.1	86,855,687	87,021,358	rs13066793	bc
chr3	q12.1	99,700,887	99,764,670	rs506186	bc
chr3	q23	141,352,866	141,441,725	rs7625643	bc
chr3	q26.31	172,538,202	172,578,973	rs78105464	bc
chr4	p14	38,747,178	38,910,312	rs10034903	bc

chr4	q13.3	74,416,747	74,623,258	rs10034692, rs7659874	md, size
chr4	q21.23	83,389,275	83,554,957	rs6854739	bc
chr4	q24	105,132,996	105,281,535	rs17617028	bc
chr4	q34.1	174,896,233	174,952,162	rs7664956, rs62334412	bc, bc
chr5	p15.33	1,271,917	1,310,593	rs10069690, rs2736107, rs150804576	bc, bc, bc
chr5	p15.1	16,188,640	16,287,055	rs12652713	bc
chr5	p13.3	32,537,171	32,590,371	rs2012709	bc
chr5	p12	44,397,125	44,586,991	rs5867671	bc
chr5	p12	44,703,112	44,875,406	rs10941679	bc
chr5	p12	44,988,392	45,332,406	rs190443933	bc
chr5	q11.2	56,368,461	56,384,203	rs7730210	bc
chr5	q11.2	56,616,049	56,617,468	rs984113	bc
chr5	q11.2	56,706,716	56,929,322	rs17432750, rs112497245, rs62355902	bc, bc, bc
chr5	q11.2	58,884,543	58,890,189	rs1353747	credvar
chr5	q11.2	58,945,289	58,946,391	rs1353747	credvar
chr5	q11.2	58,958,409	59,100,468	rs537267133	bc
chr5	q11.2	59,167,307	59,295,618	rs10472097	bc
chr5	q14.3	91,281,831	91,510,439	rs1964292	bc
chr5	q23.2	123,116,631	123,121,697	rs186749	md
chr5	q31.1	133,030,891	133,114,405	rs571173399	bc
chr5	q33.3	158,753,989	158,838,109	rs31864	bc
chr5	q35.1	170,092,648	170,169,740	rs56722914	bc
chr6	p23	13,637,145	13,759,235	rs405447	bc
chr6	p22.3	16,398,285	16,404,352	rs3819405	bc
chr6	p22.3	20,530,968	20,538,981	rs2223621	credvar
chr6	p22.3	20,541,106	20,547,026	rs2223621	credvar
chr6	p22.3	20,548,446	20,558,030	rs2223621	credvar
chr6	p22.3	20,559,446	20,563,898	rs2223621	credvar
chr6	p22.3	20,569,568	20,574,665	rs2223621	credvar
chr6	p22.3	20,582,401	20,593,538	rs2223621	credvar
chr6	p22.3	20,609,906	20,613,229	rs2223621	credvar
chr6	p22.3	20,619,886	20,623,052	rs2223621	credvar
chr6	p22.3	20,624,256	20,729,197	rs2328531	bc
chr6	q14.1	81,461,149	81,638,989	rs7763102	bc
chr6	q22.33, q23.1	129,998,090	130,086,670	rs6569648	bc
chr6	q25.1	149,257,466	149,266,042	rs35409891	bc
chr6	q25.1	151,589,730	151,670,924	rs12665607, rs7763637, rs9397437, rs12173562	md, bc, size, bc
chr6	q25.1	151,685,747	151,707,091	rs851984	bc
chr6	q25.1	151,727,606	151,758,234	rs9918437	bc
chr6	q25.1	151,975,135	152,098,859	rs79388591	bc
chr6	q25.1, q25.2	152,099,001	152,120,175	rs34133739	bc
chr7	p15.3	21,869,830	21,907,871	rs7971	bc
chr7	q21.3	94,451,081	94,676,307	rs1879854	bc
chr7	q22.1	101,842,925	101,916,979	rs7796917	bc
chr7	q32.3	130,971,457	130,996,248	rs12706954, rs6973318	bc, bc
chr7	q34	140,240,223	140,259,499	rs2003526	bc
chr7	q35	144,347,571	144,445,683	rs62485509	bc
chr8	p23.3	204,492	293,429	rs34810249	bc
chr8	p12	29,619,279	29,675,153	rs7465364	bc

chr8	p11.23	36,799,886	37,004,434	rs7816345, rs10110651, rs4286946	md, size, bc
chr8	q21.13	75,315,712	75,353,618	rs11373454	bc
chr8	q21.13	75,373,804	75,534,319	rs17303163, rs199660865	bc, bc
chr8	q22.3	101,316,631	101,418,947	rs7813150	bc
chr8	q23.1	105,303,441	105,387,155	rs150957507	bc
chr8	q23.3	115,999,849	116,006,774	rs10641009	bc
chr8	q23.3	116,195,633	116,283,097	rs13267382	bc
chr8	q24.13	123,542,045	123,606,784	rs58847541	bc
chr8	q24.13	123,716,436	123,756,860	rs4871411	bc
chr8	q24.21	127,195,124	127,214,360	rs35961416	bc
chr8	q24.21	127,320,752	127,383,897	rs419018, rs10096351	bc, bc
chr8	q24.21	128,149,254	128,214,145	rs7017073	bc
chr9	p21.3	21,943,142	22,009,008	rs539723051	bc
chr9	q31.2	107,537,236	107,549,472	rs60037937	bc
chr9	q31.2	108,009,572	108,076,487	rs10816625, rs13294895	bc, bc
chr9	q31.2	108,113,175	108,178,237	rs659713	bc
chr9	q33.1	116,533,231	116,536,115	rs13294352	bc
chr10	p14	9,019,827	9,148,294	rs7081544	bc
chr10	p12.31	21,528,645	21,557,059	rs7098100	bc
chr10	p12.31	22,168,196	22,240,695	rs138026227	bc
chr10	q21.2	62,317,602	62,437,849	rs3081227	size
chr10	q21.2	62,471,493	62,547,464	rs10509168, rs10995190, rs10995201	md, md, bc
chr10	q22.3	79,073,584	79,103,820	rs754416	bc
chr10	q22.3	79,121,808	79,138,110	rs10762851	bc
chr10	q22.3	79,455,970	79,562,928	rs61862474	bc
chr10	q25.2	112,985,904	113,054,955	rs71973726	bc
chr10	q25.3	113,361,602	113,408,515	rs12250948	bc
chr10	q26.12	121,325,863	121,338,584	rs9421409	bc
chr10	q26.13	121,570,266	121,631,551	rs2981578, rs45631563	bc, bc
chr11	p15.5	770,427	836,267	rs200835870	bc
chr11	p15.5	1,857,563	1,935,751	rs620315, rs3817198	bc, md
chr11	q13.1	65,774,554	65,793,538	rs3903072	credvar
chr11	q13.1	65,798,901	65,925,150	rs548082010	bc
chr11	q13.3	69,327,306	69,366,736	rs7102705	size
chr11	q13.3	69,507,543	69,567,476	rs78540526, rs671888, rs657686	bc, bc, bc
chr11	q22.3	108,173,573	108,432,844	rs368848598	bc
chr11	q24.3	129,572,788	129,611,689	rs745382	bc
chr12	p13.1	14,251,148	14,274,270	rs12422552	bc
chr12	p11.22	27,962,551	28,035,234	rs1838564, rs7297051	size, bc
chr12	q22	95,632,234	95,637,441	rs17356907	bc and size
chr12	q23.2	102,561,354	102,724,505	rs703556	md
chr12	q24.21	114,393,161	114,449,373	rs1265507	md
chr12	q24.21	115,141,991	115,146,732	rs35422	bc
chr12	q24.21	115,344,206	115,401,753	rs1882155, rs1353783	bc, bc
chr12	q24.21	115,757,861	115,782,683	rs11067765	bc
chr12	q24.23, q24.31	120,286,655	120,431,938	rs184486140	bc
chr13	q13.1	32,294,044	32,604,944	rs11571833	bc
chr13	q22.1	73,230,442	73,249,660	rs6562760	credvar
chr13	q22.1	73,376,293	73,398,300	rs11382527	bc
chr14	q13.3	36,604,331	36,668,665	rs12881240	bc

chr14	q13.3	36,756,140	36,811,672	rs848088	bc
chr14	q24.1	68,092,570	68,194,273	rs2478777	bc
chr14	q24.1	68,503,626	68,582,264	rs35378451	bc
chr14	q32.11	91,256,349	91,286,036	rs11341843	bc
chr14	q32.12	92,601,665	92,654,161	rs78440108	bc
chr15	q26.1	90,957,821	91,020,105	rs12594752	bc
chr16	p13.3	3,910,175	3,949,732	rs8063564	bc
chr16	p13.3	4,083,593	4,144,046	rs6500580	bc
chr16	q12.1, q12.2	52,502,401	52,605,448	rs4784227	bc
chr16	q12.2	53,758,081	53,844,605	rs55872725, rs9925952	bc, bc
chr16	q12.2	54,639,330	54,650,631	rs28539243	bc
chr16	q23.2	80,599,526	80,623,207	rs9938021	bc
chr16	q24.1, q24.2	86,984,890	87,063,432	rs4066743	bc
chr17	q11.2	30,724,981	30,938,028	rs199661266	bc
chr17	q11.2	30,940,956	30,951,569	rs146699004	credvar
chr17	q21.2	42,223,278	42,230,109	rs148509105	bc
chr17	q22	54,993,770	55,152,622	rs244321	bc
chr17	q25.3	79,793,499	79,833,708	rs2587505	bc
chr18	q11.2	26,620,331	26,627,813	rs4800749	bc
chr18	q11.2	26,740,173	26,761,720	rs527616	bc
chr18	q11.2	26,916,731	26,939,734	rs2307561	bc
chr18	q12.1	27,827,264	27,905,184	rs12970390	bc
chr18	q12.1	32,302,773	32,463,099	rs117618124	bc
chr18	q12.3	44,785,036	44,844,654	rs78955132	bc
chr18	q12.3	45,294,396	45,342,455	rs9952980	bc
chr19	p13.13	12,947,227	12,997,354	rs78269692	bc
chr19	p13.13	13,046,791	13,169,227	rs78269692	bc
chr19	p13.11	17,096,922	17,113,752	rs67397200	bc
chr19	p13.11	17,255,659	17,334,293	rs67397200	bc
chr19	p13.11	18,396,204	18,527,717	rs8105994	bc
chr19	q12	29,773,484	29,838,538	rs17513613	bc
chr19	q13.31	43,773,782	43,813,880	rs56344893	bc
chr19	q13.32	45,644,440	45,692,573	rs74174203	bc
chr19	q13.32	45,693,244	45,708,153	rs71338792	credvar
chr20	p12.3	5,966,527	5,971,824	rs16991615	bc
chr21	q11.2, q21.1	14,969,432	15,069,566	rs2822999	bc
chr21	q21.1	15,185,021	15,214,135	rs2403907	bc
chr22	q12.1	28,364,287	28,713,175	rs186430430	bc
chr22	q12.1	28,734,534	28,743,311	rs5997389	bc
chr22	q13.1	38,103,381	38,227,758	rs5995543	bc
chr22	q13.1	38,231,733	38,257,143	rs7289126	md
chr22	q13.1	38,756,058	38,873,647	rs9619765	bc
chr22	q13.1	38,960,254	38,971,081	esv3647749	bc
chr22	q13.1, q13.2	40,333,496	40,674,752	rs17001868, rs66987842, rs5995875	md, bc, size
chr22	q13.2	41,206,521	41,260,434	rs73161324	credvar
chr22	q13.2	41,315,240	41,329,250	rs73161324	credvar
chr22	q13.2	41,641,941	41,898,479	rs8137282	bc