THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Computational analysis of DNA repair pathways in breast cancer

Chao Liu

Bachelor's and PhD in Microbiology

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

Institute for Molecular Bioscience

Abstract

The integrity of the human genome is constantly challenged by a variety of endogenous and exogenous factors such as ultraviolet radiation and cigarette smoke. To deal with these threats, five major DNA repair pathways, which are principally defined by the type of lesions they repair, have evolved. Defects in these repair pathways predispose individuals to a wide variety of cancers, and at the same time can be therapeutically exploited to target tumours with defective DNA repair. Dysregulation of these repair pathways are also frequently observed in cancer, presenting both opportunities and challenges for cancer therapy: downregulated repair pathways sensitise tumours to DNA-damaging therapies, while upregulated repair pathways cause resistance to these therapies. The primary aim of this thesis is to obtain a comprehensive and in-depth understanding of the mechanisms and roles of these major DNA repair pathways in the context of breast cancer. This will be beneficial for predicting response to radiation and chemotherapy, and for developing novel targeted therapies in this common type of malignancy.

By careful literature search and consulting a domain expert, the research presented in this thesis started with a manual curation of six DNA repair pathways, including the five major repair pathways and the Fanconi anaemia pathway that is closely associated with breast cancer susceptibility. Six comprehensive pathway figures were generated, each for one repair pathway, describing in total 195 genes and 138 reactions with direct relevance to DNA repair. Moreover, to facilitate a deep understanding of the repair mechanisms, a detailed description for each reaction was given, importantly including the literature references used for curating the reaction. This curation work enables a mechanistic understanding of how cells respond to DNA damage, and provides a solid foundation for the subsequent computational analyses.

In the second study of this PhD research, I performed a personalised pathway analysis to investigate the status of homologous recombination (HR) pathway dysregulation in individual sporadic breast tumours, its association with HR repair deficiency and its impact on tumour characteristics. Specifically, using the expression values of the HR genes curated in the previous study, I calculated an HR score for each tumour that quantifies the extent of HR pathway dysregulation in that tumour. Based on that score, I observed a great diversity in HR dysregulation between and within gene expression-based breast cancer subtypes. And by comparing to two published HR-defect signatures, I found HR pathway dysregulation reflects HR repair deficiency. Furthermore, I uncovered a novel association between HR pathway dysregulation and chromosomal instability (CIN): tumours with more-dysregulated HR tend to have higher CIN. Although CIN has long been considered to be a

hallmark of most solid tumours, with recent studies highlighting its importance in tumour evolution and drug resistance, the molecular basis of CIN in sporadic cancer remains poorly understood. The novel association revealed in this study implies that HR pathway dysregulation is an important determinant of CIN in sporadic breast cancer, and thus helps pinpoint the causative factors of CIN in breast and other sporadic cancers.

The third study is a multi-omics data analysis that aimed to dissect the underlying mechanisms of DNA repair dysregulation in breast cancer. Specifically, I assessed the contributions of DNA copy number alteration (CNA), DNA methylation at gene promoter regions, and expression changes of transcriptional factors (TFs) to the differential expression of individual DNA repair genes in breast tumour versus normal samples. These gene-specific results were summarised at pathway level to estimate whether different DNA repair pathways are influenced in distinct manner. In particular, TFs potentially associated with each differentially expressed DNA repair gene were identified using a regularised linear regression-based statistical framework developed in this study. The results suggest that CNA and expression changes of TFs are major factors for DNA repair dysregulation in breast cancer, and that a limited number of TFs with multiple targets in various repair pathway may exert a global impact on repair dysregulation in this malignancy. This study thus provides new insights into the underlying mechanisms of DNA repair dysregulation in breast cancer. These insights improve our understanding of the molecular basis of the DNA repair biomarkers identified thus far, and have potential to inform future biomarker discovery in this common cancer type.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my research higher degree candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis.

Publications during candidature

Peer-reviewed papers

Liu, C., Srihari, S., Lê Cao, KA., Chenevix-Trench, G., Simpson, P.T., Ragan, M.A. and Khanna, K.K. (2014) A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. Nucleic Acids Research, 42, 6106–6127.

Srihari, S., Madhamshettiwar, P.B., Song, S., ***Liu, C.***, Simpson, P.T., Khanna, K.K. and Ragan, M.A. (2014) Complex-based analysis of dysregulated cellular processes in cancer. BMC Systems Biology, 8 (Suppl 4), S1.

Liu, C., Srihari, S., Lal S., Gautier, B., Simpson, P.T., Khanna, K.K., Ragan, M.A. and Lê Cao, KA. (2015) Personalised Pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. Molecular Oncology, 10, 179-193.

Publications included in this thesis

Liu, C., Srihari, S., Lê Cao, KA., Chenevix-Trench, G., Simpson, P.T., Ragan, M.A. and Khanna, K.K. (2014) A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. Nucleic Acids Research, 42, 6106–6127. – incorporated as Chapter 2.

Contributor	**Statement of contribution**
Author Liu C (Candidate)	Designed experiments (50%)
	Analysis and interpretation of data (60%)
	Wrote the paper (40%)
Author Srihari S	Designed experiments (10%)
	Analysis and interpretation of data (25%)
	Wrote the paper (30%)
Author Lê Cao KA	Designed experiments (0%)
	Analysis and interpretation of data (5%)
	Wrote the paper (0%)
Author Chenevix-Trench G	Designed experiments (0%)
	Analysis and interpretation of data (5%)
	Wrote the paper (0%)
Author Simpson PT	Designed experiments (5%)
	Analysis and interpretation of data (5%)
	Wrote the paper (5%)
Author Ragan MA	Designed experiments (20%)
	Analysis and interpretation of data (0%)
	Wrote the paper (10%)
Author Khanna KK	Designed experiments (15%)
	Analysis and interpretation of data (0%)
	Wrote the paper (15%)

Liu, C., Srihari, S., Lal S., Gautier, B., Simpson, P.T., Khanna, K.K., Ragan, M.A. and Lê Cao, KA. (2015) Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. Molecular Oncology, 10, 179-193 – incorporated as Chapter 3.

Contributor	**Statement of contribution**
Author Liu C (Candidate)	Designed experiments (60%)
	Analysis and interpretation of data (85%)
	Wrote the paper (50%)
Author Srihari S	Designed experiments (0%)
	Analysis and interpretation of data (0%)
	Wrote the paper (15%)
Author Lal S	Designed experiments (0%)
	Analysis and interpretation of data (10%)
	Wrote the paper (0%)
Author Gautier B	Designed experiments (0%)
	Analysis and interpretation of data (5%)
	Wrote the paper (0%)
Author Simpson PT	Designed experiments (5%)
	Analysis and interpretation of data (0%)
	Wrote the paper (5%)
Author Khanna KK	Designed experiments (5%)
	Analysis and interpretation of data (0%)
	Wrote the paper (10%)
Author Ragan MA	Designed experiments (15%)
	Analysis and interpretation of data (0%)
	Wrote the paper (10%)
Author Lê Cao KA	Designed experiments (15%)
	Analysis and interpretation of data (0%)
	Wrote the paper (10%)

Contributions by others to the thesis

No contributions by others.

Statement of parts of the thesis submitted to qualify for the award of another degree

None.

Acknowledgements

Looking back the past four years, it has been an amazing journey for me. And along the way there are so many wonderful people, including mentors, peers and family, who have made the completion of this journey possible. The following is just a humble record of some of these individuals, to whom there are no words that can fully express my gratitude.

First, a big "thank you" to my principal advisor, Prof. Mark Ragan for his mentoring and guidance throughout my candidature. He introduced me to the Systems Biology world and opened a new chapter in my career. He has been guiding me through this journey and watching me grow. This thesis would not be possible without him.

I would like to express my gratitude to my co-advisor, Dr. Kim-Anh Lê Cao for her guidance and support throughout this journey, especially those in my most difficult days. As a biostatistician, she has taught me a lot about statistical analysis, and she has always been able to find time from her tight schedule to meet with me on a regular basis. I would also like to acknowledge my co-advisor, Prof. Kum Kum Khanna for selflessly sharing with me her expertise in DNA repair and breast cancer. Her help enables me to stand on a giant's shoulders and see much more than what I could on my own. I am also so grateful to my co-advisor, Dr. Peter Simpson for his guidance in terms of breast cancer biology, his warm attitude throughout this long journey, and all the inspiring ideas and suggestions that he has shared with me.

I would also like to extend my warm thanks to members of Ragan Group, especially Dr. Sriganesh Srihari and Dr. Cheong-Xin Chan, for the valuable discussions we have had over the past four years. I am also deeply grateful for Dr. Amanda Carozzi, Ms. Olga Chaourova and Ms. Cody Mudgway for their administrative support throughout my candidature.

Last but not least, I would like to thank my Mum, Dad and my wife Jessie. Their unconditional love and company are always my source of courage and strength.

Keywords

personalised pathway analysis, multi-omics data analysis, integrative analysis, penalised linear regression, DNA repair pathway, DNA repair dysregulation, chromosomal instability, breast cancer

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060102, Bioinformatics, 50%

ANZSRC code: 060114, Systems Biology, 30%

ANZSRC code: 060405 Gene Expression, 20%

Fields of Research (FoR) Classification

FoR code: 0601, Biochemistry and Cell Biology, 80%

FoR code: 0604, Genetics, 20%

Table of Contents

List of Figures & Tables

All figures and tables included within this thesis are contained within published or submitted manuscripts (Chapters Two, Three, Four and Five) and are numbered and referenced therein.

List of Abbreviations

A-NHEJ	Alternative Non-Homologous End Joining
AML	Acute Myeloid Leukemia
ATM	Ataxia Telangiectasia Mutated
ATR	Ataxia Telangiectasia and Rad3 related
AUC	Area Under the Curve
BER	Base Excision Repair
C-NHEJ	Canonical Non-Homologous End Joining
CIN	Chromosomal Instability
CNA	Copy Number Alteration
DDR	DNA Damage Response
DE	Differentially Expressed
DM	DNA Methylation
DNA-PKcs	DNA-dependent Protein Kinase catalytic subunit
DRPFS	DNA Repair Pathway-Focused Score
DSB	Double-Strand Break
DSBR	DSB Repair
ER	Estrogen Receptor
FA	Fanconi Anaemia
FCS	Functional Class Scoring
FDR	False Discovery Rate
GO	Gene Ontology
GSA	Gene Set Analysis
GSEA	Gene Set Enrichment Analysis
HDACi	Histone Deacetylase inhibitors
HJ	Holliday Junctions
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
IDL	Insertion/Deletion Loops
iPAS	individualised Pathway Aberrance Score

IPL	Integrated Pathway Level
KEGG	Kyoto Encyclopedia of Genes and Genomes
LST	Large-Scale Transition
METABRIC	Molecular Taxonomy of Breast Cancer International Consortium
MMEJ	Microhomology-Mediated End Joining
MMR	Mismatch Repair
MSE	Mean Squared Error
NER	Nucleotide Excision Repair
NHEJ	Non-Homologous End Joining
ORA	Over-Representation Analysis
PCA	Principal Component Analysis
PDS	Pathway Deregulation Score
PR	Progesterone Receptors
PT	Pathway Topology
SFE	Selected Functional Events
SL	Synthetic Lethality
SSB	Single-Strand Break
TCGA	The Cancer Genome Atlas
TF	Transcriptional Factors
TNBC	Triple Negative Breast Cancer
TSS	Transcription Start Site
UV	Ultraviolet Radiation

Chapter One: Introduction

1.1 Breast cancer classification

Breast cancer is one of the most common malignancies worldwide, with more than 1,300,000 cases diagnosed each year, along with 450,000 deaths (1). It is also the leading cause of disease burden in Australia, with 14,560 cases reported nationwide in 2012 (2).

Breast cancer is a highly heterogeneous disease that consists of a spectrum of subtypes with distinct morphological features, variable clinical outcomes and different responses to different therapies (2, 3). To dissect this heterogeneity, several classification schemes have been proposed (Figure 1), which are described as follows:

1.1.1 Resulted from germ-line mutation or sporadic genetic change

Based on whether the disease is caused by germ-line mutation or sporadic genetic change, breast cancer can be divided into two categories: familial (inherited) and sporadic. Familial breast cancer represents ~7% of breast cancer cases, and is mainly due to inherited mutation in DNA repair genes (4, 5). For example, ~30% cases of familial breast patients are carriers of mutations in either *BRCA1* or *BRCA2*, two essential genes for DNA repair (6).

Sporadic breast cancer, which arises primarily as a result of genetic alterations acquired during a person's lifetime, is the dominant form of breast cancer. Although familial breast tumours also show genetic aberrations other than gene mutations, sporadic breast tumours exhibit much-more complicated patterns of genetic changes including focal mutations, epigenetic alterations, small insertions and/or deletions of DNA segments, and even large-scale chromosomal abnormalities (7, 8). Interestingly, many sporadic genetic changes are also associated with DNA repair genes (9, 10), highlighting the intimate relationship between breast cancer susceptibility and DNA repair abnormality.

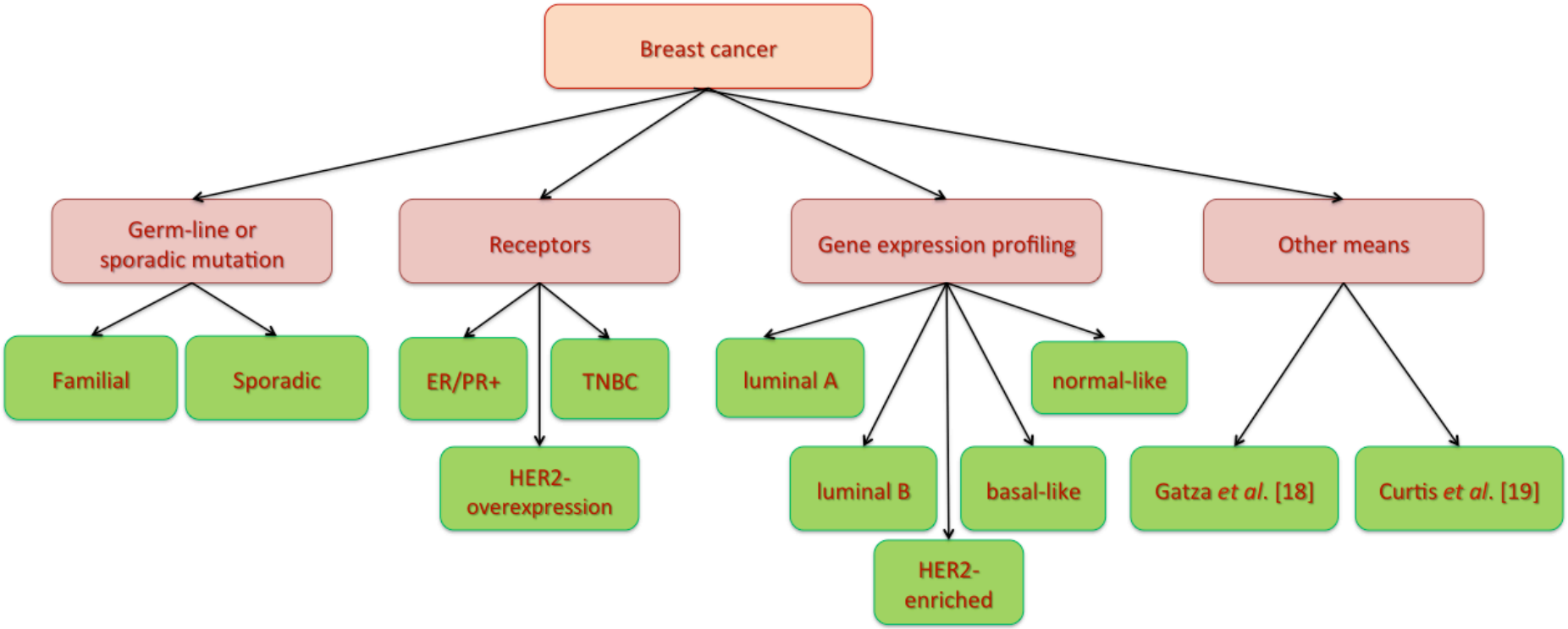


Figure 1 Breast cancer subtypes based on different classification schemes

1.1.2 Hormone and growth factor receptor-based

A second classification of breast cancer, which is based on two hormone receptors [Estrogen Receptor (ER) and Progesterone Receptors (PR)] and one epidermal growth factor receptor [Human Epidermal growth factor Receptor 2 (HER2)], has been routinely used in the clinical practice of breast cancer management for about two decades (9-11). This classification scheme divides breast cancer into three groups: hormone receptor positive, hormone receptor negative with HER2 over-expression, and triple negative (i.e., breast tumours that do express none of these three receptors). Each of these three subtypes is summarised as follows:

1.1.2.1 Hormone receptor positive

This subtype comprises ~70% of breast cancer cases (12). Cancer cells of this subtype express at least one of the two hormone receptors (i.e., ER and PR), and are dependent on hormone signals to promote their growth. As a result, tumours of this subtype are likely to respond to anti-hormone drugs such as tamoxifen, and hence are not considered as refractory when a combination of surgery, chemotherapy (or radiation) and hormonal therapy is used. However, if tumours of this subtype recur, they tend to evolve into hormone-insensitive forms by acquiring resistance to anti-hormone drugs (9, 10).

1.1.2.2 Hormone receptor negative with HER2 over-expression

Tumours in this subtype correspond to ~15% of breast cancer patients (13). The cells in these tumours express neither ER nor PR, but instead exhibit amplification and overexpression of HER2, a tyrosine kinase transmembrane receptor that promotes cell growth. Tumours of this subtype are in general more aggressive than the hormone-positive tumours, but some of them respond well to anti-HER2 agents such as trastuzumab (9, 10).

1.1.2.3 Triple negative

This subtype accounts for ~15% of breast tumours (14), and is often referred to as Triple Negative Breast Cancer (TNBC) as cancer cells of this subtype do not express any of the three aforesaid receptors. In most cases, TNBC tumours are much more aggressive than tumours that express at least one of the three receptors. As of now, the pathogenesis of TNBC remains poorly understood, and cytotoxic agents such as anthracycline are the major therapeutic option for this subtype. As a consequence, TNBC patients are characterised by poor prognosis including high recurrence rate and low five-year survival (15, 16).

The study of TNBC has gained growing interest in recent years. There is increasing evidence showing that *BRCA1* inactivation may play an important role in the development of this

subtype, as most breast tumours with *BRCA1* mutation belong to TNBC (14). Besides, recent developments of novel target therapies based on the concept of *Synthetic Lethality* (SL) have brought hope for better treatment in this subtype. More detail about SL-based therapies is given below in Section 3.2 of this chapter, and in Chapter Two.

1.1.3 Gene expression profile-based

The development of high-throughput technologies, in particular microarrays, has made it possible to refine breast cancer classification based on whole-genome molecular profiles that better reflect the underlying biology of tumours. As pioneers in this regard, Perou et al. (17) first classified breast tumours into four intrinsic subtypes (luminal, HER2-enriched, basal-like and normal-like) by using hierarchical clustering analysis of microarray data. These four intrinsic subtypes are of clinical relevance and have been confirmed by a number of subsequent studies [for example, see (18-21)]. Some of the subsequent studies also revealed that the luminal subtype can be further divided into two subtypes with distinct prognoses: luminal A and luminal B [see (22) for a review]. To date, these five gene expression-based subtypes (also known as the PAM50 subtypes) have been widely used in research. A brief summary about each of them is given below.

1.1.3.1 Luminal A

Luminal A is the most prevalent subtype, representing ~50%-60% of all breast cancer cases (22). Tumours in this subtype are characterised by increased expression of ER and/or PR-related genes, and decreased expression of genes involved in cell proliferation. Patients of this subtype generally have a lower relapse rate and longer survival time compared with other subtypes (9, 22).

1.1.3.2 Luminal B

Luminal B tumours account for ~10%-20% of breast cancers. Similar to luminal A tumours, luminal B tumours normally exhibit elevated expression of ER and/or PR-associated genes. The difference between these two subtypes is that luminal B tumours also display increased expression of proliferation-related genes such as *MKI67* and *CCNB1*. As a consequence, luminal B tumours are more aggressive and have worse prognosis than luminal A tumours (9, 22).

1.1.3.3 HER2-enriched

About 15%-20% of breast cancer tumours correspond to the HER2-enriched subtype. This subtype is characterised by high expression of *HER2* as well as some other genes whose genomic locations are close to *HER2*, and of genes involved in cell proliferation. In addition, ~40% HER2-enriched tumours harbour p53 mutations. Clinically, although HER2-enriched tumours are highly

proliferative, patients of this subtype are getting better prognosis with the development of anti-HER2 therapies (9, 22).

1.1.3.4 Basal-like

Tumours of basal-like subtype constitute ~10%-20% of all breast carcinomas. Tumours in this subtype normally express one or more of the basal cytokeratins (e.g. *CK5*, *CK14* and *CK17*) and usually display an elevated *p53* mutation rate. Familial breast cancers with germ-line *BRCA1* mutations are mostly clustered into this subtype (9, 22).

Compared with other subtypes, basal-like tumours tend to occur at an early age, with larger tumour size and higher histological grade. Consequently, although basal tumours in general are more sensitive to chemotherapy, they usually have a worse prognosis than do other subtypes (9, 22).

In clinical practice, the terms basal-like and TNBC are often considered as synonymous, but according to microarray-based measurements up to 30% genes display distinct expression patters between these two categories. Therefore, basal-like and TNBC should be regarded as distinct but intersecting classes (15).

1.1.3.5 Normal-like

Normal-like breast tumours are rare, accounting for about 5%–10% of the diagnosed cases. The clinical significance of subtype has yet to be determined due to its rarity. As normal-like tumours show similar expression pattern to normal breast tissues, some researchers even argue that this subtype might correspond to tumour samples contaminated by normal breast tissues during sample preparation and/or microarray experiment (22).

1.1.4 Recent advances in breast cancer classification

Despite the critical roles that the above-mentioned classification schemes have played in breast cancer research and treatment, substantial variation still exists within each subtype. Several studies have been conducted in recent years to tackle this issue. As one major effort, Gatza et al. (23) suggested 17 subgroups of breast cancer based on pathway activity patterns using gene expression data. Specifically, subgroups 11 and 17 correspond to luminal A tumours; subgroups 3, 4, 6, 9 and 16 represent luminal B tumours; subgroups 7 and 10 are composed of HER2-enriched tumours; subgroups 2, 5, and 8 are basal-like tumours; and the other subgroups are a mixture of varied PAM50 subtypes. This new breast cancer taxonomy provides a functional interpretation of each subgroup and therefore can be particularly useful in guiding therapeutic choices and patient stratification for testing new drugs (23).

As another major effort, by integrating DNA copy number and gene expression data, Curtis et al. classified breast cancer samples into ten integrative clusters (8). Of these clusters, clusters 3, 7, 8 are primarily luminal A tumours; clusters 1, 6, 9 contain luminal B tumours; cluster 10 corresponds to basal-like tumours; and the other clusters are composed of tumours from mixed PAM50 subtypes. This genome-driven integrated classification represents an important advance in understanding the genomic diversity of breast cancer, and therefore is expected to have profound implications for the rationale development of tailored breast cancer therapy (8, 24).

Some other studies also aimed to refine the commonly used breast cancer classifications. For example, Lehmann et al. identified seven sub-categories within TNBC (25) according to gene expression profiles, and Ciriello et al. defined four major subtypes within luminal A tumours through an integrative analysis of DNA copy number data and mutation data (26). However, it is likely that heterogeneity still exists within these newly established subtypes. To achieve the aim of personalised medicine, in the future each tumour needs be analysed individually.

1.2 Overview of DNA repair pathways

The cellular DNA repair machinery is crucial for maintaining the integrity of human genome, which is constantly challenged by a variety of endogenous and exogenous factors, including ultraviolet radiation (UV), chemical carcinogens and oxidative by-products from normal cellular respiration. This repair machinery can be generally divided into five distinct but functional interlinked pathways: homologous recombination (HR), non-homologous end joining (NHEJ), nucleotide excision repair (NER), base excision repair (BER) and mismatch repair (MMR) (Figure 2). Defects in these repair pathways predispose individuals to a variety of cancers, and the cellular status of these repair pathways can be a key determinant of cancer outcome following chemotherapy, radiotherapy and some targeted therapies. Below is a brief overview for each of these pathways, emphasising their clinical relevance. A detailed description regarding the mechanistic aspect of each pathway can be found in Chapter Two.

1.2.1 The HR Pathway

The HR pathway represents an error-free mechanism mainly for the repair of double-strand breaks (DSBs) during DNA replication, and is thus vital for the high-fidelity transmission of genetic information across generations. Many breast cancer susceptibility genes are involved in this pathway, including *BRCA1* and *BRCA2* as mentioned above, and also *ATM*, *PALB2*, *BRIP1* and *RAD51L1* (27). Defects in HR also predispose to many other cancers, including ovarian, prostate and pancreatic cancer (28-30).

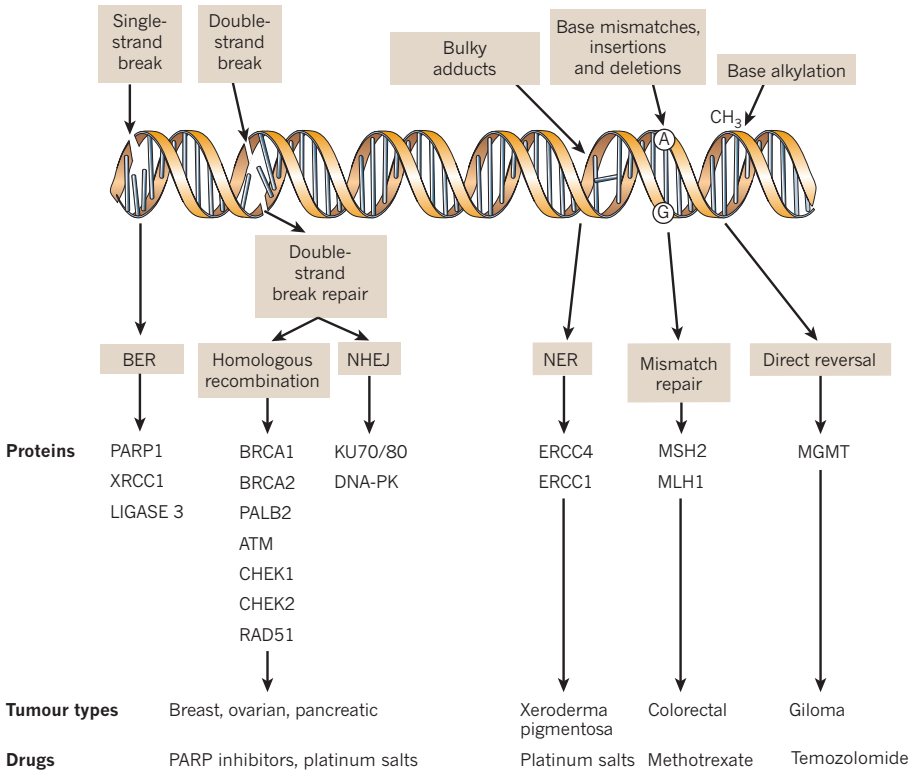


Figure 2 A panoply of DNA repair mechanisms maintains genomic stability (from Lord et al., 2012). Direct reversal repair is not considered as a pathway as it involves only one enzyme.

HR is the most-complicated DNA repair pathway, containing the largest number of repair proteins. Many of the these proteins have been exploited intensively as drug targets or biomarkers because defects in HR sensitise tumours to DSB-inducing agents such as ionising radiation and topoisomerase I poisons, and platinum-based agents such as cisplatin and carboplatin (30). Clinically, the most-exciting discovery in the field of HR in recent years is the identification of the SL relationship between HR deficiency and poly (ADP-ribose) polymerase (PARP) inhibition, which is described in more detail below.

1.2.2 The NHEJ Pathway

NHEJ is another major pathway to repair DSB. However, unlike HR, which employs a sophisticated mechanism to ensure accuracy of the repair, NHEJ ligates the two broken ends of a DSB in a direct way, which is faster than the HR repair but tends to be more error-prone. The choice between NHEJ and HR is primarily dependent on the cell cycle stage – HR normally occurs during S and G2 phases of the cell cycle as it requires a homologous sister chromatid to serve as a template for repair, whereas NHEJ predominantly functions in the G0 and G1 phases, although it can also come into play in other cell-cycle phases especially when the HR repair becomes compromised (31).

The genetic alterations in key NEHJ genes have been linked to cancer preposition. For example, mutations in *KU70* have been associated with susceptibility to breast cancer, colorectal cancer and lung cancer, and lung cancer can also be caused by epigenetic silencing of *KU80* (30). In addition, increased activity of NHEJ has been proposed as a main source of genomic rearrangements observed in various cancers (32).

1.2.3 The NER Pathway

The NER pathway deals with DNA double helix-distorting damage, such as that induced by UV light or tobacco smoke. Defective NER predisposes individuals to different cancer-prone syndromes, including xeroderma pigmentosum, cockayne syndrome, and trichothiodystrophy, each of which is characterised by a high risk of skin cancer (33). Genetic alterations in key NER genes may also give rise to other cancers. For example, *ERCC1* methylation and the polymorphisms observed in *XPA* and *XPC* have been implicated in the genesis of lung cancer and bladder cancer, respectively (30, 34). NER deficiency confers sensitivity to various chemotherapeutic agents including cisplatin, mitomycin and nitrogen mustard (33).

1.2.4 The BER Pathway

The BER pathway is primarily responsible for removing small, non-helix-distorting base lesions caused by oxidation, alkylation or deamination, which often induces single-strand breaks (SSBs). Genetic defects in essential BER genes such as *MUTYH*, *OGG1* and *MTH1* are primarily coupled with excess risk of colorectal cancer (35). In addition, the polymorphisms observed in *OGG1* has been implicated in lung cancer susceptibility (36).

Compromised BER activity renders tumour cells sensitive to ionising radiation and alkylating agents such as methyl methanesulphonate and temozolomide. Especially, alkylating agents represent a primary class of front-line chemotherapeutic drugs, whose efficacy is largely influenced by the cellular status of BER and MMR pathways (37).

1.2.5 The MMR Pathway

The MMR pathway is the major mechanism for the repair of base-base mismatches and insertion/deletion loops (IDL) formed during DNA replication. In MMR-deficient tumour cells, mutation rates are up to 1,000-fold greater than normal cells (38). Germ-line mutations in central MMR genes including *MSH2*, *MSH6*, *PMS2* and *MLH1* can cause Lynch syndrome, an inherited disorder associated with an elevated lifetime risk for colorectal cancer, endometrial cancer, ovarian cancer and stomach cancer (39). MMR deficiency can cause hypersensitivity of tumours to alkylating agents, as mentioned above (37). Moreover, the cellular status of MMR can also affect

the resistance of tumours to ionising radiation (40) and chemotherapeutic agents such as methotrexate, anthracycline and taxane (41).

In addition to these five major repair pathways mentioned above, in Chapter Two I also provided a detailed mechanistic description for the Fanconi anaemia (FA) pathway as this pathway is closely associated with breast cancer susceptibility. In brief, FA is a rare genetic disorder characterised by bone marrow failure, susceptibility to breast and other cancers, and hypersensitivity to DNA inter-strand crosslink agents such as cisplatin. This disorder is caused by mutations in a cluster of DNA repair-related genes that function in the FA pathway (42, 43).

1.3 DNA repair pathways and breast cancer therapy

Although defects in distinct DNA repair pathways are connected with different cancer susceptibilities, as described above, it appears that the occurrence of breast cancer has a particularly close relationship with DNA repair deficiency. For instance, in an epidemiological study comparing 285 women with breast cancer and 539 women without breast cancer, Matta et al. revealed a significant correlation between reduced overall DNA repair capacity and elevated breast cancer risk (44). The biological mechanism behind this phenomenon has not been explained convincingly, but one suggestion is that cells in mammary tissue have a higher rate of proliferation, apoptosis and differentiation compared with cells in most of other human tissues (10).

DNA repair pathways have important implications for radiation therapy and chemotherapy, as described above. Moreover, the research of DNA repair pathways in breast cancer has gained more attention in recent years due to the encouraging progress in the developments of SL-based targeted therapies. In the following section, I briefly discuss the relevance of various DNA repair pathways in the context of different breast cancer therapies.

1.3.1 DNA repair pathways and radiotherapy/chemotherapy response

Current breast cancer therapy typically involves surgery in combination with various DNA damaging agents including ionising radiation, platinum-based drugs, anthracycline and taxane. Ionising radiation induces SSBs, DSBs and oxidised bases. The primary component of platinum-based drugs is an alkylating compound that causes intra- and inter-strand crosslinks. Anthracyclines mainly function as topoisomerase II inhibitors, giving rise to DSBs. Taxane agents are mitotic inhibitors that disrupt the process of mitotic cell division (9, 10).

Breast tumours show subtype-specific response to DNA-damaging agents, although in most cases the underlying mechanisms are not completely understood and substantial exceptions still exist. For example, basal-like tumours normally exhibit hypersensitivity to platinum-based drugs such as cisplatin, which is generally believed to be the consequence of HR deficiency (45). As

another example, HER2-enriched tumours frequently exhibit sensitivity to anthracycline and taxane, but not to platinum-based drugs (9, 46, 47). The underlying mechanisms of these observations are not clear so far, and possibly multiple mechanisms are involved, including NER, HR and NHEJ (9). Luminal tumours, which include luminal A and luminal B, are more diverse in terms of drug response, and it has been proposed that the altered activities of HR, NHEJ and/or BER might be involved in this diversity (9).

In recent years, small-molecule inhibitors that specifically target each DNA repair pathway have been under active development. The major rationale behind this development is that DNA repair pathways with increased activities can cause resistance to DNA-damaging therapies. Therefore, in addition to damaging the DNA, targeting these pathways by specific inhibitors is likely to achieve enhanced therapeutic effects (9, 30). As an example in breast cancer, Huang et al. demonstrated that inhibition of RAD51 by a small molecule termed B02 results in diminished HR, which in turn leads to significantly increased tumour sensitivity to cisplatin (48). Developments in this field are likely to have substantial impact on breast cancer therapy in the near future.

1.3.2 SL-based targeted therapy

The concept of SL first arose from genetic studies in *Drosophila* and is now often used to refer to a type of genetic interaction in which the co-occurrence of two genetic events results in cell death, while the occurrence of one event is still compatible with cell viability (49-51). In the field of oncology, it has been shown that tumour cells can become 'addicted' to compensatory DNA repair pathways for survival if they already acquired one defective DNA repair pathway at their origin (49). This addiction can be therapeutically exploited based on the concept of SL, which represents a promising direction for developing targeted therapy that can effectively kill cancer cells while at the same time spare normal cells.

A good example of applying SL in breast cancer therapy is the development of PARP inhibitors for tumours that are deficient in HR due to mutations in *BRCA1* or *BRCA2* (52, 53). Although these tumours exhibit extraordinary sensitivity to PARP inhibitors at concentrations that are safe for normal tissues (52, 53), the exact mechanism underlying this observation remains somewhat contentious. It was suggested that the inhibition of PARP results in compromised BER repair, leading to accumulation of SSBs that will be converted into DSBs during DNA replication. In normal cells, these resultant DSBs can be repaired by HR, but in cancer cells that are defective in HR, these DSBs will accumulate and eventually induce cell death (54). This explanation, however, has been challenged by alternative models. In particular, Patel et al. (55) showed that the NHEJ pathway is the major contributor to the cytotoxicity generated by PARP inhibitors in HR-deficient

cells. A detailed description regarding the recent advances in this field is given in the publication (51) incorporated in Chapter Two.

To summarise the biological part of this thesis introduction: a major obstacle in current breast cancer management is the extensive heterogeneity observed among breast cancer patients. Successfully dissecting this heterogeneity, ideally at the individual level, should lead to enormous improvements in treatment effectiveness. The cellular status of DNA repair pathways is one of the decisive factors underlying the sensitivity and resistance of individual tumours to various cancer therapies; however, much is still unknown regarding the mechanisms and status of these highly complex pathways. This thesis thus aims to perform a computational analysis of these repair pathways in the context of breast cancer, making use of the cutting-edge bioinformatics tools that have emerged in recent years.

In the remaining sections, I summarise the development of pathway analysis approaches in the field of bioinformatics, with emphasis on the pros and cons associated with each type of methods, and on the challenges that remain.

1.4 Overview of pathway analysis approaches

1.4.1 Introduction

The advent of high-throughput whole-genome profiling techniques, such as microarray and RNA-Seq, has greatly enhanced biological research by allowing genome-wide measurements of molecular features in a single experiment (56, 57). Univariate single gene-based analysis (e.g. t- or F- test) of the high-throughput data typically yields a list of differentially expressed (DE) genes between two different phenotypes (56, 58). This list, however, is often inadequate in providing functional insights into the underlying mechanisms that drive the phenotypic distinction. To overcome this limitation, new methods termed pathway analysis have been developed, which shift the analysis from gene level to pathway level (56, 59-66).

A biological pathway corresponds to a set of proteins that participate in the same biological process in a cell. The term "pathway analysis" has been widely used in the literature. For example, it has been applied to describe kinetic simulation of pathways, flux-balance analysis of steady-state pathways and inference of novel pathways from high-throughput data (56, 61, 62). In this thesis, I use pathway analysis to describe approaches that statistically test one or multiple pathways for significant association with a phenotype. These approaches are sometimes also called "knowledge-based pathway analysis" [for example in (61)] to emphasise that they make use of prior biological knowledge about the pathways being studied. In many other articles, these approaches are referred to as gene set enrichment analysis (GSEA) or gene set analysis (GSA), especially when the sets of

genes being analysed do not correspond to biological pathways [e.g., gene sets defined by gene ontology (GO) terms] (59, 62, 67, 68).

It needs to be pointed out that the biological meanings of "pathway" and "gene set" are substantially different. These two terms seem to be interchangeable in some literature, especially when enrichment analyses are used to identify pathways enriched for differentially expressed genes. However, equating a pathway with a gene set disregards knowledge of gene interactions within the pathway. Furthermore, a pathway can be regarded as a functional biological unit, while the genes in a gene set do not necessarily participate in the same biological process (e.g., signature gene sets derived from genome-scale gene expression studies), or represent only parts of pathways (e.g., gene sets including only pathway component genes differentially expressed in tumours). Therefore, genes sets may provide incomplete information or include irrelevant genes when they are used to investigate, for instance, which biological processes are altered between two conditions. In light of this, I use pathways rather than gene sets throughout this thesis.

The application of pathway analysis for interpreting high-throughput data has exploded over the past decade for the following reasons:

a) From a biological point of view, a list of pathways identified as altered between two phenotypes has more explanatory power for explaining the phenotypic difference than does a list of DE genes (56, 57).

b) From a statistical point of view, grouping tens of thousands of genes into hundreds of pathways is advantageous as the dimensionality of the data is reduced. As a result, the number of statistical hypotheses that need to be tested is also much reduced (59).

c) Pathway analysis is capable of detecting weak but coordinated expression changes of genes within the same pathway (56, 57). These changes may have a significant impact on the phenotypic difference being studied but can be easily missed by assessing each gene separately. For example, Mootha et al. found no single DE gene between Type II diabetes positive and negative patients, but identified a set of genes that show subtly but coordinated expression changes in diabetes positive patients. This set of genes is involved in the oxidative phosphorylation pathway, suggesting a role of this pathway in the development of diabetes (69).

1.4.2 Classification of pathway analysis approaches

A large number of pathway analysis approaches has been developed so far. According to Khatri et al. (61), these approaches can be classified into three categories: over-representation analysis (ORA) approaches, functional class scoring (FCS) approaches, and pathway topology (PT)-based approaches (Figure 3).

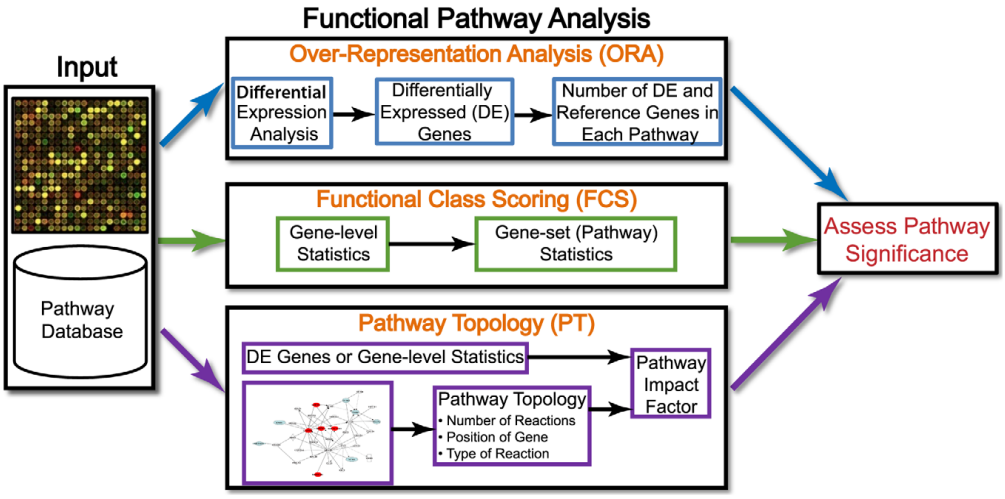


Figure 3 Three categories of pathway analysis approaches (from Khatri et al., 2012).

1.4.2.1 ORA approaches

ORA approaches represent early efforts in the field of pathway analysis. In general, these approaches use a contingency table to test whether a pathway is overrepresented by a predefined list of DE genes. The univariate statistical tests adopted in these approaches include the hypergeometric test, the chi-square test and the binomial test. Despite the popularity of ORA approaches in GO-based tools such as BiNGO (70), GOstat (71) and GOEAS (72), they share two significant drawbacks:

a) They all use an arbitrary threshold (e.g., genes with fold-change ≥ 2 and/or p-values ≤ 0.05) to obtain a list of DE genes. This may lead to severe information loss as a substantial portion of the genes is discarded, especially those that are marginally less significant (e.g., fold-change = 1.999 and/or p-value = 0.051) (61, 65, 66).

b) The statistical tests adopted by these approaches consider only the number of DE genes, and ignore any quantitative values associated with the genes (e.g., fold-change, p-value etc.). This may also result in significant information loss (61, 65, 66).

c) These statistical tests all assume that the component genes of the same pathway are independent of each other, resulting in information loss. More-advanced statistical analyses that account for inter-gene correlations, such as ROAST (73) and CAMERA (74), have been developed to tackle this issue.

1.4.2.2 FCS approaches

FCS approaches represent the mainstream methods for the current pathway-based analysis of high-throughput data, which include gene set enrichment analysis (GSEA) (75), the most prominent method in the field. An important hypothesis underlying FCS approaches is that although

DE genes can have most significant effects on the activity of a pathway, the statistic of a given pathway should be based on the statistic of all genes that function in that pathway. Specifically, there are two reasons that make FCS approaches superior than ORA methods (56, 61, 68):

a) FCS approaches do not require an arbitrary threshold for separating expression data into significant and non-significant parts and thus make full use of all experimental data.

b) FCS approaches use gene-specific molecular measurements to detect coordinated changes in the expression of genes in the same pathway.

1.4.2.3 PT-based approaches

A significant drawback associated with FCS approaches is that they ignore the topology of a pathway (i.e, the relative positions of component genes in a pathway as well as the number and types of interactions between the genes). As a consequence, the output of FCS approaches will remain unchanged even if a pathway is redrawn with new connections between the component genes (61).

The PT-based approaches, which utilise pathway topology to calculate pathway statistics, are thought to be able to overcome the drawback of the FCS methods (61). For now, only a small number of PT-based approaches have been reported, which include ScorePAGE (76), Pathway-Express (77), SPIA (78) and NetGSA (79). Some of these approaches are not implemented in software or a package (e.g., ScorePAGE and NetGSA), while others may have functional restrictions (e.g., Pathway-Express is available only as a Web server). Moreover, at the moment PT-based approaches have difficulties in modelling large complex pathways (e.g. the DNA repair pathways) due to the complexity of interactions within these pathways. Those difficulties as well as the non user-friendly features currently make PT-based approaches less than ideal for pathway analysis.

1.4.3 Steps of FCS approaches

As FCS approaches play a dominant role in the current pathway analysis-based studies, in this section I summarise the common steps of FCS approaches and discuss challenges that are associated with each step (Figure 4).

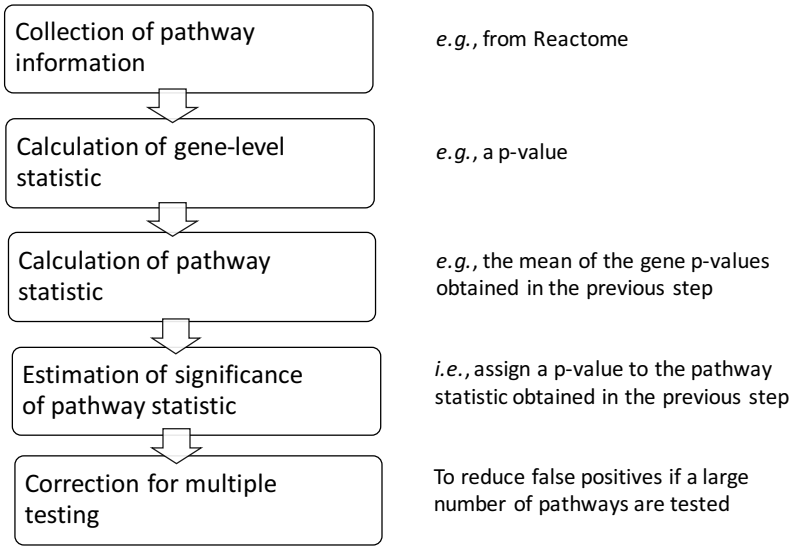


Figure 4 Steps of Functional Class Scoring (FCS) approaches for pathway analysis.

1.4.3.1 Collection of pathway information

The first step in applying FCS approaches is to retrieve biological knowledge for pathways that are to be investigated. This step is not specific to FCS approaches, but may have significant impact on the quality of their outputs. Commonly-used public pathway databases include Reactome (80), KEGG (81), MetaCyc (82), PID (83), PantherDB (84), WikiPathways (http://wikipathways.org/), STKE (http://stke.sciencemag.org/cm/), BioCarta (http://www. biocarta.com) and ResNet (http://www.ariadnegenomics.com/). Detailed description of these pathway databases is beyond the scope of this introduction. Interested readers are referred to excellent reviews (85-87) that provide nice summaries of commonly used pathway databases. Particularly, the recent work by Chowdhury et al. summarises and compares the properties of 24 pathway databases, including their in-built technical features and their respective merits and demerits (85).

There are two major problems associated with the current public pathway databases. First, as different pathway databases adopt distinct strategies for pathway inclusion and curation, these databases differ greatly in terms of content quality. In particular, some databases such as ResNet were developed by electronic curation (i.e., they use data-mining algorithms to infer functional relationships between genes). Although these inferred annotations are useful for hypothesis generation, their accuracy is usually in doubt, and this has limited the usefulness of such databases for high-quality pathway analysis. By contrast, many other databases such as Reactome, KEGG and MetaCyc employ manual curation for their development and maintenance. Although relatively time-consuming, manual curation can lead to more accurate and up-to-date results, which is particularly useful for pathways that are complicated and/or poorly annotated.

Second, even for pathway databases that are manually curated, similarly named pathways across different databases may exhibit substantial differences in constitution, whereas differently named pathways across databases may exhibit considerable overlap. For example, Altman et al. performed a systematic comparison between KEGG and MetaCyc databases and found many KEGG pathways are much larger in size compared with their MetaCyc counterparts (87). As another example, Chowdhury et al. demonstrated significant data heterogeneity between 24 pathway databases, including Reactome, KEGG and MetaCyc (85).

1.4.3.2 Calculation of gene-level statistic

The second step is to compute a gene-level statistic for each gene in the same pathway. This step is relatively simple and often corresponds to, for example, a single gene-based differential expression analysis in a microarray experiment. A gene-level statistic can be a t-statistic, a p-value, a signal-to-noise ratio (mean to standard deviation ratio), a log-likelihood ratio, a fold change, or a Wilcoxon rank sum statistic (59, 62). An interesting point to note is that the choice of gene-level statistic seems to have negligible effect on identifying significant pathways (59).

Different technologies for measuring gene expression levels, such as RNA-seq and microarray, might also affect the calculation of gene-level statistic. RNA-seq is thought to be superior to microarray in detecting low abundance transcripts and novel isoforms, and is also free of issues inherent to microarray, such as probe redundancy, cross-hybridisation and non-specific hybridisation. Besides, to obtain a gene-level measurement, RNA-seq usually adopts the union of the transcripts that represent the same gene while microarray often considers only the most-abundant probe of the gene. However, recent large-scale comparisons showed that these two techniques produce highly-correlated results (Spearman correlation coefficient of 0.8) (88) or that the choice between them is not a significant factor affecting the final results (89). Together with the facts that microarray is more cost-effective, and that a great wealth of microarray-based expression data has already been accumulated, the results from these two gene expression profiling techniques are complementary to each other.

1.4.3.3 Calculation of pathway statistic

The next step is to calculate a pathway-level statistic by aggregating the gene-level statistic of all component genes in a pathway. A pathway-level statistic can be univariate, such as the Kolmogorov-Smirnov statistic generated by GSEA (75), the maxmean (90) and the Wilcoxon rank sum (91), which disregards interdependencies (e.g., correlations) among genes within a pathway; it can also be multivariate, such as the Hotelling's T^2 (92, 93) and the statistic generated by GlobalANCOVA (94), which accounts for interdependencies among the pathway component genes.

Excellent reviews (56, 59, 61-63) have summarised and compared the commonly-used pathway-level statistics.

It is tempting to speculate that a multivariate statistic may have a higher statistical power than a univariate statistic because the former takes more information into consideration. However, Glazko et al. showed that although multivariate statistics exhibited higher statistical power than univariate statistic on simulated data, when applied to real biological data, univariate statistics displayed more power at stringent cutoffs (p-value ≤ 0.001), and equal power at less-stringent cutoffs (p-value ≤ 0.05) [(95); reviewed in (61)]. Considering that multivariate statistics are generally much more computationally expensive, univariate statistics still remain the common choice in current pathway analysis (59, 61).

Regardless of the differences between univariate and multivariate statistic, the power of a pathway statistic is largely determined by the proportion of DE genes within the pathway, the size of the pathway and the strength of correlations between the expression of the component genes of a pathway (61).

1.4.3.4 Estimation of significance of pathway statistics

A further step is to estimate the statistical significance of the pathway statistic calculated in the previous step. The result of this step strongly depends on the choice of null hypothesis. As first defined by Tian et al. (96) and further described by Ackermann and Strimmer (59), there are two types of null hypothesises, termed *competitive null hypothesis* and *self-contained null hypothesis*:

a) The competitive null hypothesis permutes gene labels for each pathway, and compares the genes in a pathway with the genes that are not in the pathway. The rationale in this hypothesis is that a significantly altered pathway between two phenotypes should be distinguishable from equal sized pathways that are composed of randomly selected genes (59, 96).

b) The self-contained null hypothesis permutes phenotype labels for each sample and compares the genes in a pathway with themselves. The rationale in this hypothesis is that the association of a pathway with a phenotype change should be distinguishable by randomly shuffling phenotype labels (59, 96).

There is no consensus in the literature regarding which hypothesis is better. Many researchers favour the self-contained null hypothesis, as the results directly address the question of finding pathways whose expression change correlates with the phenotype change (56, 66, 97). Others prefer the competitive null hypothesis, as sample permutation takes much more computational time, and the results based on the two hypotheses are similar to each other (68, 98).

1.4.3.5 Correction for multiple testing

Although a p-value is considered to be an appropriate measure of statistical significance when one single pathway is tested, if a large number of pathways are tested, there can be many false positives among pathways with small p-values. Multiple hypotheses correction is therefore needed to correct the p-values of multiple pathways obtained in the previous step (59, 62).

Bonferroni-related approaches represent a straightforward way for multiple test correction, and simply multiply the p-values obtained in each test by the total number of hypotheses (i.e. the total number of pathways being tested). Despite their simplicity to understand and implement, these approaches do not account for pathway overlap and dependence, and therefore may result in a large number of false negatives (62, 99).

In recent years, False Discovery Rate (FDR) methods have become widely used for multiple testing correction in pathway analysis (61, 62). Several variations of FDR with different assumptions of the underlying data distribution have been proposed, including the Benjamini-Yekutieli (BY) correction (100), the Benjamini-Hochberg (BH) correction (101), positive FDR (pFDR) (102) and significant analysis of microarray (SAM) FDR (103). Kim et al. compared these FDR methods using random correlation matrices, and found that BH was the most robust method (104).

1.4.4 Recent advances in pathway analysis approaches

Despite the great number of pathway analysis methods that have been developed, and the significant role these methods have played in interpreting genome-wide molecular measurements, the vast majority of these methods focus on identifying altered pathways between two groups (e.g. cancer versus normal), and thus cannot provide pathway information for individuals. This issue is becoming a critical concern since cancer is a heterogeneous disease. Dissecting this heterogeneity will be crucial for understanding the underlying mechanisms and disease status of each tumour, and for developing tailed therapies that target specific pathways. In the following part, I discuss several tools that have been recently developed for individualised pathway analysis.

Ahn et al. (105) extended several existing ORA- and FCS-based pathway analysis methods to generate individualised pathway aberrance score (iPAS) for each tumour. Overall, the steps for generating iPAS are similar to the procedures described in Section 4.3, with the exception that the pathway statistic is calculated by comparing each individual tumour sample with many accumulated normal samples (i.e., normal samples from patients with the same type of disease). The authors applied iPAS to lung adenocarcinoma and colon cancer samples respectively, and showed that this score can provide biologically and clinically relevant representation for the individual tumours. In particular, they found the iPAS based on the “amino acid synthesis and interconversion” pathway

can be used to identify lung adenocarcinoma samples from unknown samples [Area Under the Curve (AUC) = 0.982], which cannot be achieved by using the conventional ORA or FCS methods (105).

Wang et al. (106) developed individPath to detect pathways with significantly disrupted coordination of gene expression for individual disease samples. This was achieved by performing intra-pathway gene pair comparisons between each disease and accumulated normal samples. One key advantage of individPath is its robustness to batch effects as relative expression orderings of genes within samples are insensitive to normalization methods. The effectiveness of individPath for personalised pathway analysis was demonstrated by the identification of a prognostic intra-pathway gene pair signature for early-stage lung adenocarcinoma, and an intra-pathway gene pair signature that is predictive of relapse-free survival of ER^+ breast cancer patients after tamoxifen treatment (106).

The Pathifier method proposed by Drier et al. (107) quantifies a pathway's dysregulation in individual tumours by calculating a pathway deregulation score (PDS) separately for each pathway in every sample, with accumulated normal samples serving as a reference. Specifically, to calculate the PDS for pathway *P*, Pathifier first fits a principal curve that captures the maximal variability of the expression of all genes in pathway *P* in both tumour and normal samples, and then projects all samples onto that curve; a tumour's PDS for that pathway is defined as the distance along the curve from the projection of the tumour to the centroid of the projections of normal samples (107). Through the analysis of three colorectal cancer datasets and two glioblastoma datasets, the authors showed that PDSs can consistently dissect heterogeneity of pathway dysregulation in a sensible, valid and clinically useful manner (107). Moreover, Pathifier has been successfully applied to provide a pathway-based classification of breast cancer (108), and when combined with Cox regression and L1 penalised estimation has achieved better prognosis prediction compared with gene-based models (109).

The pathway recognition algorithm using data integration on genomic models (PARADIGM) developed by Vaske et al. (110) represents another prominent tool for personalised pathway analysis. Unlike the aforementioned methods, a pathway component considered by PARADIGM is not limited to mRNA values; it can be any molecular measurement of a gene (copy number, mRNA or protein level), a protein complex, a gene family or even an abstract process such as "apoptosis". For each sample and each pathway component, PARADIGM builds a probabilistic graphical model to generate a single summary, called integrated pathway level (IPL), to indicate the status of a pathway component in a given sample relative to a benchmark (e.g. as measurements in normal tissues). The IPLs belonging to the same pathway are then summarised to represent the activity of that pathway in a given tumour.

PARADIGM has been successfully applied in various cancer studies (110-113). Compared with the methods mentioned above, PARADIGM might perform better with multi-omics data, as it takes into consideration the relationships between different pathway components. However, it might not perform well with single-omics data and/or when relationships between different pathway components are not well understood (107). Moreover, PARADIGM was not available until recently, which had limited its wide usage.

One common problem associated with the above-mentioned personalised pathway analysis methods is that instead of comparing tumour and normal samples from the same individual, they make use of accumulated normal samples. Although this is a good strategy for now as matched normal samples are often unavailable, the results of these methods can be inaccurate due to interpersonal heterogeneity. Furthermore, intra-tumour heterogeneity has gained increased attention in recent years (114), and with the development of new single-cell sequencing technologies, it has been suggested that whole-genome sequencing of a number of separate cells in a single tumour will become a necessity in the near future (115). Accordingly, pathway analysis tools that account for intra-tumour heterogeneity need to be developed.

To summarise the bioinformatics part of this Introduction, the pathway information retrieved from various pathway databases differs in quality, which may exert notable influence on pathway analysis result. Besides, although FCS methods remain the mainstream approaches for pathway analysis, and ORA methods were thought as the future (61), personalised pathway methods are emerging as promising tools to rise up the challenge posed by the tremendous heterogeneity observed in cancer. I would anticipate more personalised pathway analysis methods to be developed in the coming years.

To address some of the issues raised in this introduction, three studies were conducted and are presented in this thesis (chapters Two, Three and Four). The specific issue addressed is summarised at the beginning of each chapter. In general, these studies are driven by various biological questions regarding DNA repair and breast cancer, and involve applying state-of-the-art bioinformatics methods to the increasing amount of genomic data to formulate knowledge that is of biological relevance and clinical implication. In this sense, this thesis serves to bridge the gap between biological findings and bioinformatics developments.

REFERENCES

1. Koboldt,D.C., Fulton,R.S., McLellan,M.D., Schmidt,H., Kalicki-Veizer,J., McMichael,J.F., Fulton,L.L., Dooling,D.J., Ding,L., Mardis,E.R., *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

2. Australian Institute of Health and Welfare (2012) Cancer in Australia: an overview 2012. Cancer

series no. 74.

3. Viale,G. (2012) The current state of breast cancer classification. *Annals of Oncology*, **23**, 207–210.

4. Claus,E.B., Schildkraut,J.M., Thompson,W.D. and Risch,N.J. (1996) The genetic attributable risk of breast and ovarian cancer. *Cancer*, **77**, 2318-2324.

5. kConFab Investigators, Waddell,N., Arnold,J., Cocciardi,S., da Silva,L., Marsh,A., Riley,J., Johnstone,C.N., Orloff,M., Assie,G., *et al.* (2009) Subtypes of familial breast tumours revealed by expression and copy number profiling. *Breast Cancer Res Treat*, **123**, 661–677.

6. Shih,H.A., Couch,F.J., Nathanson,K.L., Blackwood,M.A., Rebbeck,T.R., Armstrong,K.A., Calzone,K., Stopfer,J., Seal,S., Stratton,M.R., *et al.* (2002) BRCA1 and BRCA2 mutation frequency in women evaluated in a breast cancer risk evaluation clinic. *Journal of Clinical Oncology*, **20**, 994–999.

7. Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

8. Curtis,C., Shah,S.P., Chin,S.-F., Turashvili,G., Rueda,O.M., Dunning,M.J., Speed,D., Lynch,A.G., Samarajiwa,S., Yuan,Y., *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346-352.

9. Li,S.-X., Sjolund,A., Harris,L. and Sweasy,J.B. (2010) DNA repair and personalized breast cancer therapy. *Environmental and Molecular Mutagenesis*, **51**, 897–908.

10. Davis,J.D. (2011) DNA damage and breast cancer. *World Journal of Clinical Oncology*, **2**, 329.

11. Polyak,K. and Vogt,P.K. (2012) From the Cover: Progress in breast cancer research. *Proceedings of the National Academy of Sciences*, **109**, 2715–2717.

12. Schnitt,S.J. (2010) Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. *Modern Pathology*, **23 Suppl 2**, S60–4.

13. Ravdin,P.M. and Chamness,G.C. (1995) The c-erbB-2 proto-oncogene as a prognostic and predictive marker in breast cancer: a paradigm for the development of other macromolecular markers--a review. *Gene*, **159**, 19–27.

14. Gelmon,K., Dent,R., Mackey,J.R., Laing,K., McLeod,D. and Verma,S. (2012) Targeting triple-negative breast cancer: optimising therapeutic outcomes. *Annals of Oncology*, **23**, 2223–2234.

15. Arnedos,M., Bihan,C., Delaloge,S. and Andre,F. (2012) Triple-negative breast cancer: are we making headway at least? *Therapeutic Advances in Medical Oncology*, **4**, 195–210.

16. Elias,A.D. (2010) Triple-Negative Breast Cancer. *American Journal of Clinical Oncology*, **33**, 637–645.

17. Perou,C.M., Sørlie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., Rees,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

18. Gruvberger,S., Ringnér,M., Chen,Y., Panavally,S., Saal,L.H., Borg A, Fernö,M., Peterson,C. and Meltzer,P.S. (2001) Estrogen receptor status in breast cancer is associated with remarkably

distinct gene expression patterns. *Cancer Research*, **61**, 5979–5984.

19. West,M., Blanchette,C., Dressman,H., Huang,E., Ishida,S., Spang,R., Zuzan,H., Olson,J.A., Marks,J.R. and Nevins,J.R. (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, **98**, 11462–11467.

20. Sørlie,T., Perou,C.M., Tibshirani,R., Aas,T., Geisler,S., Johnsen,H., Hastie,T., Eisen,M.B., van de Rijn,M., Jeffrey,S.S., *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences*, **98**, 10869–10874.

21. van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A.M., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T., *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

22. Eroles,P., Bosch,A., Pérez-Fidalgo,J.A. and Lluch,A. (2012) Molecular biology in breast cancer: Intrinsic subtypes and signaling pathways. *Cancer Treatment Reviews*, **38**, 698–707.

23. Gatza,M.L., Lucas,J.E., Barry,W.T., Kim,J.W., Wang,Q., Crawford,M.D., Datto,M.B., Kelley,M., Mathey-Prevot,B., Potti,A., *et al.* (2010) A pathway-based classification of human breast cancer. *Proceedings of the National Academy of Sciences*, **107**, 6994–6999.

24. Dawson,S.-J., Rueda,O.M., Aparicio,S. and Caldas,C. (2013) A new genome-driven integrated classification of breast cancer and its implications. The *EMBO Journal*, **32**, 617–628.

25. Lehmann,B.D., Bauer,J.A., Chen,X., Sanders,M.E., Chakravarthy,A.B., Shyr,Y. and Pietenpol,J.A. (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *Journal of Clinical Investigation*, **121**, 2750–2767.

26. Ciriello,G., Sinha,R., Hoadley,K.A., Jacobsen,A.S., Reva,B., Perou,C.M., Sander,C. and Schultz,N. (2013) The molecular diversity of Luminal A breast tumors. *Breast Cancer Research Treatment*, **141**, 409–420.

27. Mavaddat,N., Antoniou,A.C., Easton,D.F. and García-Closas,M. (2010) Genetic susceptibility to breast cancer. *Molecular Oncology*, **4**, 174–191.

28. Cerbinskaite,A., Mukhopadhyay,A., Plummer,E.R., Curtin,N.J. and Edmondson,R.J. (2012) Defective homologous recombination in human cancers. *Cancer Treatment Reviews*, **38**, 89–100.

29. Moynahan,M.E. and Jasin,M. (2010) Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nature Reviews Molecular Cell Biology*, **11**, 196–207.

30. Curtin,N.J. (2012) DNA repair dysregulation from cancer driver to therapeutic target. *Nature Reviews Cancer*, **12**, 801–817.

31. Liu,C., Srihari,S., Lal,S., Gautier,B., Simpson,P.T., Khanna,K.K., Ragan,M.A. and Lê Cao,K.-A. (2015) Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. *Molecular Oncology*, **10**, 179-193.

32. Bunting,S.F. and Nussenzweig,A. (2013) End-joining, translocations and cancer. *Nature*

Reviews Cancer, **13**, 443–454.

33. Leibeling,D., Laspe,P. and Emmert,S. (2006) Nucleotide excision repair and cancer. *Journal of Molecular Histology*, **37**, 225–238.

34. Sertic,S., Pizzi,S., Lazzaro,F., Plevani,P. and Muzi-Falconi,M. (2012) NER and DDR: Classical music with new instruments. *Cell Cycle*, **11**, 668–674.

35. Farrington,S.M., Tenesa,A., Barnetson,R., Wiltshire,A., Prendergast,J., Porteous,M., Campbell,H. and Dunlop,M.G. (2005) Germline susceptibility to colorectal cancer due to base-excision repair gene defects. *American Journal of Human Genetics*, **77**, 112–119.

36. Dianov,G.L. (2011) Base excision repair targets for cancer therapy. *American Journal of Human Genetics*, **1**, 845–851.

37. Fu,D., Calvo,J.A. and Samson,L.D. (2012) Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nature Reviews Cancer*, **12**, 104–120.

38. las Alas,de,M.M., Aebi,S., Fink,D., Howell,S.B. and Los,G. (1997) Loss of DNA mismatch repair: effects on the rate of mutation to drug resistance. *Journal of the National Cancer Institute*, **89**, 1537–1541.

39. Peltomäki,P. (2001) Deficient DNA mismatch repair: a common etiologic factor for colon cancer. *Human Molecular Genetics*, **10**, 735–740.

40. Martin,L.M., Marples,B., Coffey,M., Lawler,M., Lynch,T.H., Hollywood,D. and Marignol,L. (2010) DNA mismatch repair and the DNA damage response to ionizing radiation: making sense of apparently conflicting data. *Cancer Treatment Reviews*, **36**, 518–527.

41. Rivera,E. (2010) Implications of anthracycline-resistant and taxane-resistant metastatic breast cancer and new therapeutic options. *Breast Journal*, **16**, 252–263.

42. Moldovan,G.-L. and D'Andrea,A.D. (2009) How the Fanconi Anemia Pathway Guards the Genome. *Annual Review of Genetics*, **43**, 223–249.

43. D'Andrea,A.D. (2010) Susceptibility pathways in Fanconi's anemia and breast cancer. *New England Journal of Medicine*, **362**, 1909–1919.

44. Matta,J., Echenique,M., Negron,E., Morales,L., Vargas,W., Gaetan,F.S., Lizardi,E.R., Torres,A., Rosado,J.O., Bolaños,G., *et al.* (2012) The association of DNA Repair with breast cancer risk in women. A comparative observational study. *BMC Cancer*, **12**, 490.

45. Byrski,T., Huzarski,T., Dent,R., Gronwald,J., Zuziak,D., Cybulski,C., Kladny,J., Gorski,B., Lubinski,J. and Narod,S.A. (2009) Response to neoadjuvant therapy with cisplatin in BRCA1-positive breast cancer patients. *Breast Cancer Resesearch Treatment*, **115**, 359–363.

46. Gennari,A., Sormani,M.P., Pronzato,P., Puntoni,M., Colozza,M., Pfeffer,U. and Bruzzi,P. (2008) HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials. *Journal of the National Cancer Institute*, **100**, 14–20.

47. Hayes,D.F., Thor,A.D., Dressler,L.G., Weaver,D., Edgerton,S., Cowan,D., Broadwater,G., Goldstein,L.J., Martino,S., Ingle,J.N., *et al.* (2007) HER2 and response to paclitaxel in node-positive breast cancer. *New England Journal of Medicine*, **357**, 1496–1506.

48. Huang,F. and Mazin,A.V. (2014) A small molecule inhibitor of human RAD51 potentiates breast cancer cell killing by therapeutic agents in mouse xenografts. *PLoS ONE*, **9**, e100993.

49. Shaheen,M., Allen,C., Nickoloff,J.A. and Hromas,R. (2011) Synthetic lethality: exploiting the addiction of cancer to DNA repair. *Blood*, **117**, 6074–6082.

50. Helleday,T. (2011) The underlying mechanism for the PARP and BRCA synthetic lethality: Clearing up the misunderstandings. *Molecular Oncology*, **5**, 387–393.

51. Liu,C., Srihari,S., Lê Cao,K.-A., Chenevix-Trench,G., Simpson,P.T., Ragan,M.A. and Khanna,K.K. (2014) A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. *Nucleic Acids Research*, **42**, 6106–6127.

52. Farmer,H., McCabe,N., Lord,C.J., Tutt,A.N.J., Johnson,D.A., Richardson,T.B., Santarosa,M., Dillon,K.J., Hickson,I., Knights,C., *et al.* (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**, 917–921.

53. Bryant,H.E., Schultz,N., Thomas,H.D., Parker,K.M., Flower,D., Lopez,E., Kyle,S., Meuth,M., Curtin,N.J. and Helleday,T. (2005) Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, **434**, 913–917.

54. Rouleau,M., Patel,A., Hendzel,M.J., Kaufmann,S.H. and Poirier,G.G. (2010) PARP inhibition: PARP1 and beyond. *Nature Reviews Cancer*, **10**, 293–301.

55. Patel,A.G., Sarkaria,J.N. and Kaufmann,S.H. (2011) Nonhomologous end joining drives poly(ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells. *Proceedings of the National Academy of Sciences*, **108**, 3406–3411.

56. Emmert-Streib,F. and Glazko,G.V. (2011) Pathway Analysis of Expression Data: Deciphering Functional Building Blocks of Complex Diseases. *PLoS Computational Biology*, **7**, e1002053.

57. Varadan,V., Mittal,P., Vaske,C. and Benz,S. The Integration of Biological Pathway Knowledge in Cancer Genomics: A review of existing computational approaches. *IEEE Signal Processing Magazine*, **29**, 35–50.

58. Allison,D.B., Cui,X., Page,G.P. and Sabripour,M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat Reviews Genetics*, **7**, 55–65.

59. Ackermann,M. and Strimmer,K. (2009) A general modular framework for gene set enrichment analysis. *BMC Bioinformatics*, **10**, 47.

60. Curtis,R.K., Orešič,M. and Vidal-Puig,A. (2005) Pathways to the analysis of microarray data. *Trends in Biotechnology*, **23**, 429–435.

61. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, **8**, e1002375.

62. Hung,J.-H., Yang,T.-H., Hu,Z., Weng,Z. and DeLisi,C. (2012) Gene set enrichment analysis: performance evaluation and usage guidelines. *Briefings in Bioinformatics*, **13**, 281–291.

63. Ramanan,V.K., Shen,L., Moore,J.H. and Saykin,A.J. (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in Genetics*, **28**, 323–332.

64. Copeland,N.G. and Jenkins,N.A. (2009) Deciphering the genetic landscape of cancer – from

genes to pathways. *Trends in Genetics*, **25**, 455–462.

65. Nam,D. and Kim,S.Y. (2008) Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, **9**, 189–197.

66. Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

67. Rivals,I., Personnaz,L., Taing,L. and Potier,M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

68. Väremo,L., Nielsen,J. and Nookaew,I. (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, **41**, 4378–4391.

69. Mootha,V.K., Lindgren,C.M., Eriksson,K.-F., Subramanian,A., Sihag,S., Lehar,J., Puigserver,P., Carlsson,E., Ridderstråle,M., Laurila,E., *et al.* (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**, 267–273.

70. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*, **21**, 3448–3449.

71. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.

72. Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Research*, **36**, W358–W363.

73. Wu,D., Lim,E., Vaillant,F., Asselin-Labat,M.-L., Visvader,J.E. and Smyth,G.K. (2010) ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*, **26**, 2176–2182.

74. Wu,D. and Smyth,G.K. (2012) Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, **40**, e133.

75. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.

76. Rahnenführer,J., Domingues,F.S., Maydt,J. and Lengauer,T. (2004) Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applicatioins in Genetics and Molecular Biology*, **3**, Article16.

77. Draghici,S., Khatri,P., Tarca,A.L., Amin,K., Done,A., Voichita,C., Georgescu,C. and Romero,R. (2007) A systems biology approach for pathway level analysis. *Genome Research*, **17**, 1537–1545.

78. Tarca,A.L., Draghici,S., Khatri,P., Hassan,S.S., Mittal,P., Kim,J.-S., Kim,C.J., Kusanovic,J.P. and Romero,R. (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

79. Shojaie,A. and Michailidis,G. (2009) Analysis of gene sets based on the underlying regulatory network. *Journal of Computational Biology*, **16**, 407–426.

80. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B., *et al.* (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, **39**, D691–D697.

81. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27–30.

82. Caspi,R., Altman,T., Dale,J.M., Dreher,K., Fulcher,C.A., Gilham,F., Kaipa,P., Karthikeyan,A.S., Kothari,A., Krummenacker,M., *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Research*, **38**, D473–9.

83. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research*, **37**, D674–D679.

84. Thomas,P.D. (2003) PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Research*, **13**, 2129–2141.

85. Chowdhury,S. and Sarkar,R.R. (2015) Comparison of human cell signaling pathway databases--evolution, drawbacks and challenges. *Database (Oxford)*, **2015**.

86. Cerami,E.G., Gross,B.E., Demir,E., Rodchenkov,I., Babur,Ö., Anwar,N., Schultz,N., Bader,G.D. and Sander,C. (2011) Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research*, **39**, D685–90.

87. Altman,T., Travers,M., Kothari,A., Caspi,R. and Karp,P.D. (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.

88. Guo,Y., Sheng,Q., Li,J., Ye,F., Samuels,D.C. and Shyr,Y. (2013) Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS ONE*, **8**, e71462.

89. Zhang,W., Yu,Y., Hertwig,F., Thierry-Mieg,J., Zhang,W., Thierry-Mieg,D., Wang,J., Furlanello,C., Devanarayan,V., Cheng,J., *et al.* (2015) Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biology*, **16**, 133.

90. Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Annals of Applied Statistics*, **1**, 107–129.

91. Barry,W.T., Nobel,A.B. and Wright,F.A. (2005) Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics*, **21**, 1943–1949.

92. Kong,S.W., Pu,W.T. and Park,P.J. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.

93. Lu,Y., Liu,P.Y., Xiao,P. and Deng,H.W. (2005) Hotelling's T2 multivariate profiling for detecting differential expression in microarrays. *Bioinformatics*, **21**, 3105–3113.

94. Hummel,M., Meister,R. and Mansmann,U. (2007) GlobalANCOVA: exploration and assessment of gene group effects. *Bioinformatics*, **24**, 78–85.

95. Glazko,G.V. and Emmert-Streib,F. (2009) Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets. *Bioinformatics*, **25**, 2348–2354.

96. Tian,L., Greenberg,S.A., Kong,S.W., Altschuler,J., Kohane,I.S. and Park,P.J. (2005)

Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences*, **102**, 13544–13549.

97. Fridley,B.L., Jenkins,G.D. and Biernacka,J.M. (2010) Self-Contained Gene-Set Analysis of Expression Data: An Evaluation of Existing and Novel Methods. *PLoS ONE*, **5**, e12693.

98. Li,J., Wang,L., Xu,L., Zhang,R., Huang,M., Wang,K., Xu,J., Lv,H., Shang,Z., Zhang,M., *et al.* (2012) DBGSA: a novel method of distance-based gene set analysis. *Journal of Human Genetics*, **57**, 642–653.

99. Shaffer,J.P. (1995) Multiple hypothesis testing. *Annual Review of Psychology*, **46**, 561.

100. Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statics*, **29**, 1165-1188.

101. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

102. Storey,J.D. and Tibshirani,R. (2001) Estimating false discovery rates under dependence, with applications to DNA microarrays. *Technical of Report*, No. 2001-28.

103. Dudoit,S., Shaffer,J.P. and Boldrick,J.C. (2003) Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**, 71-103.

104. Kim,K.I. and van de Wiel,M.A. (2008) Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, **9**, 114.

105. Ahn,T., Lee,E., Huh,N. and Park,T. (2014) Personalized identification of altered pathways in cancer using accumulated normal tissue data. *Bioinformatics*, **30**, i422–9.

106. Wang,H., Cai,H., Ao,L., Yan,H., Zhao,W., Qi,L., Gu,Y. and Guo,Z. (2015) Individualized identification of disease-associated pathways with disrupted coordination of gene expression. *Briefings in Bioinformatics*, **17**, 78-87.

107. Drier,Y., Sheffer,M. and Domany,E. (2013) Pathway-based personalized analysis of cancer. *Proceedings of the National Academy of Sciences*, **110**, 6388–6393.

108. Livshits,A., Git,A., Fuks,G., Caldas,C. and Domany,E. (2015) Pathway-based personalized analysis of breast cancer expression data. *Molecular Oncology*, **9**, 1471-1483.

109. Huang,S., Yee,C., Ching,T., Yu,H. and Garmire,L.X. (2014) A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. *PLoS Computational Biology*, **10**, e1003851.

110. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–45.

111. Cancer Genome Atlas Research Network, Kandoth,C., Schultz,N., Cherniack,A.D., Akbani,R., Liu,Y., Shen,H., Robertson,A.G., Pashtan,I., Shen,R., *et al.* (2013) Integrated genomic characterization of endometrial carcinoma. *Nature*, **497**, 67–73.

112. Cancer Genome Atlas Research Network (2014) Comprehensive molecular characterization of

urothelial bladder carcinoma. *Nature*, **507**, 315–322.

113. Zack,T.I., Schumacher,S.E., Carter,S.L., Cherniack,A.D., Saksena,G., Tabak,B., Lawrence,M.S., Zhsng,C.-Z., Wala,J., Mermel,C.H., *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nature Genetics*, **45**, 1134–1140.

114. Burrell,R.A., McGranahan,N., Bartek,J. and Swanton,C. (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.

115. Stephens,Z.D., Lee,S.Y., Faghri,F., Campbell,R.H., Zhai,C., Efron,M.J., Iyer,R., Schatz,M.C., Sinha,S. and Robinson,G.E. (2015) Big Data: Astronomical or Genomical? *PLoS Biology*, **13**, e1002195.

Chapter Two: Manual curation of the DNA repair pathways

The cellular DNA repair machinery is a highly complex system, involving hundreds of proteins with diverse functions. To dissect this complexity, research on DNA repair is advancing rapidly, with the number of proteins known to be involved in this machinery expanding constantly. At the time I commenced this PhD study (2012), however, the commonly used pathway databases such as KEGG and Reactome had failed to keep their DNA repair entries up-to-date to reflect these advancements (e.g., the Reactome DNA repair pathways had not been updated since 2003). This posed a primary difficulty for an accurate computational analysis of the DNA repair pathways.

To tackle this issue, I manually curated the gene content for six DNA repair pathways (i.e., the HR, NHEJ, NER, BER, MMR and FA pathways as described in Chapter One) by literature search and consulting a domain expert. In total, this curation work covered 195 genes and 138 reactions that have direct relevance in DNA repair. For each repair pathway, information on relevant genes and reactions, which was scattered over a wide range of original publications, was assembled into a comprehensive pathway diagram. Moreover, to facilitate a deep understanding of repair mechanisms, a detailed description for each reaction is given, which includes the references used for curating the reaction. This work paved the way for the computational analyses presented in the following chapters.

Results presented as a publication

The results regarding HR and NHEJ pathways were published as a review in the peer-reviewed journal *Nucleic Acids Research* in May 2014. According to Google Scholar, this publication has attracted 26 citations as of 18th January 2016. The pathway figures and reaction descriptions for all pathways are presented in Appendix 1. The original PDF version of the pathway figures, and an Excel File containing all curated DNA repair genes are deposited at UQ eSpace.

A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy

Chao Liu[1*], Sriganesh Srihari[1*], Kim-Anh Lê Cao[1], Georgia Chenevix-Trench[2], Peter T. Simpson[3], Mark A. Ragan[1] and Kum Kum Khanna[2]

[1]Institute for Molecular Bioscience, The University of Queensland, St Lucia QLD 4072, Australia

[2]QIMR-Berghofer Medical Research Institute, Herston, Brisbane, QLD 4029, Australia

[3]The University of Queensland Centre for Clinical Research, Herston, Brisbane QLD 4029, Australia

**Equal contribution*

Correspondence: Prof Kum Kum Khanna, Signal Transduction Laboratory, QIMR Berghofer Medical Research Institute, Level 9, 300 Herston Road, Brisbane, QLD 4006, Australia. Tel.: +61 7 3362 0338; Fax: +61 7 3362 0111. E-mail: KumKum.Khanna@qimrberghofer.edu.au; Prof Mark A. Ragan, Institute for Molecular Bioscience, the University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia. Tel.: +61 7 3346 2616; Fax: +61 7 3346 2101; E-mail: m.ragan@uq.edu.au

Abstract

DNA-damage response (DDR) machinery is crucial to maintain the genomic integrity of cells, by enabling effective repair of even highly lethal lesions such as DNA double-strand breaks (DSBs). Defects in specific genes acquired through mutations, copy-number changes or epigenetic silencing can alter the balance of these pathways, triggering cancerous potential in cells. Selective killing of cancer cells by sensitizing them to further DNA damage, especially by induction of DSBs, therefore requires careful modulation of DSB-repair pathways.

Here, we review the latest knowledge on the two DSB-repair pathways, homologous recombination (HR) and non-homologous end joining (NHEJ) in human, describing in detail the functions of their components and the key mechanisms contributing to the repair. Such an in-depth characterization of these pathways enables a more mechanistic understanding of how cells respond to therapies, and suggests molecules and processes that can be explored as potential therapeutic targets. One such avenue that has shown immense promise is *via* the exploitation of synthetic lethality (SL) relationships, for which the *BRCA1-PARP1* relationship is particularly notable. Here we describe how this relationship functions and the manner in which cancer cells develop therapy resistance by restoring their DSB repair potential.

1. Introduction

Most DNA-damaging chemotherapeutic agents directly or indirectly cause DNA double-strand breaks (DSBs), which are highly lethal lesions sufficient to kill cells by inactivating essential genes or, in metazoans, by triggering apoptosis (1, 2). The key to highly selective cancer therapies therefore lies in exploiting the distinctive molecular and cellular traits that sensitise only cancer cells to these agents.

Cancer is a disease of genomic instability and cancer cells differ genetically from normal cells in their ability to repair their DNA. Consequently, if these differences can be exploited to induce a high level of DNA damage, which can nonetheless be repaired in normal cells, then cancer cells can be selectively forced into DNA-damage-induced apoptosis. DNA-damage response (DDR) pathways offer molecular targets to exploit cancer-specific traits and through their precise modulation, cancer cells can be selectively sensitised to DSB-inducing drugs.

Cells have evolved an intricate assembly of interlocking mechanisms that repair DSBs efficiently or, if the damage cannot be repaired, commit the cells to apoptosis. Extensive studies mapping mutational landscapes of cancers have linked specific defects in DSB-repair pathways to 'driver' events in breast and other cancers (3, 4). It is also now established that cancer cells become drug-resistant and retain their proliferative potential by modulating their DSB-repair potential (5). Therefore in-depth characterization of DSB-repair pathways and deciphering their connection to tumorigenic activity is critical to understand the basis of cancer and develop effective therapies.

In the following section, we describe the basic mechanisms underlying DSB-repair and associated sub-pathways, from sensing of DNA damage and recruitment of early-response factors through to repair and the re-joining of DNA ends. In the subsequent section, through associating specific genes and mechanisms in these pathways to cancerous potential particularly for breast cancer, we outline how this information can be harnessed to improve cancer therapy, focusing on a promising strategy called *synthetic lethality*.

2. DNA damage response (DDR)

The detection of DSBs activates a sequence of closely linked cellular events, designated the DDR, consisting of cell-cycle checkpoint activation, chromatin modification, transcriptional changes, DNA repair, or apoptotic cell death in cases where the damage cannot be repaired [see (1, 6-8) for more details]. The principal function of this regulatory network is to maximise the likelihood that any genetic lesion incurred is faithfully repaired prior to being transmitted to progeny during DNA replication or mitotic cell division. Critical regulators of cell cycle

checkpoints include the ATM (ataxia telangiectasia mutated) and ATR (ataxia telangiectasia and Rad3 related) protein kinases, which act in concert or independently to deal with DNA damage in the cell (9, 10). A large-scale proteomics screen identified greater than 700 proteins phosphorylated by ATM and/or ATR in response to genotoxic stress, demonstrating the broad impact of DNA damage on cellular signalling (11-13). The checkpoint functions of ATR and ATM are mediated, in part, by a pair of checkpoint effector kinases termed CHK1 (checkpoint kinase 1) and CHK2 (checkpoint kinase 2) [reviewed in (14)]. Another direct target of ATM phosphorylation relevant to G1 phase cell cycle arrest is p53 (9), one of the most important tumour suppressors. Together with its key target p21, p53 plays an important role in inducing cell cycle arrest and regulating the balance between repair and survival of the cell or apoptosis [(15, 16); recently reviewed in (8)]. In addition to the classical transducers (ATM and ATR) and effector kinases (CHK1 and CHK2), stress-activated p38 SAPK (stress-activated protein kinase) and its downstream target MAPKAP-kinase 2 (Mitogen-Activated Protein Kinase-Activated Protein Kinase 2) (17, 18) and tyrosine kinases such as Abl (Abelson murine leukemia) play an important role in coordinating the DDR of higher eukaryotic cells (19, 20). Description of all the DNA damage response-induced pathways is beyond the scope of this article, and the reader is referred to several excellent reviews (1, 6, 7, 21); only salient features of DSB repair will be highlighted here.

3. DSB repair

Homologous recombination (HR) and Non-homologous end joining (NHEJ) are the two main DSB repair pathways. HR restores the original DNA sequence at DSB sites using a template sequence from a sister chromatid or a homologous chromosome to direct the error-free repair of DSBs, and is restricted to the S and G2 phases of the cell cycle. In addition to DSB repair, HR is also involved in the resolution of stalled replication forks and in the generation of genetic diversity through mitotic and meiotic recombination (22, 23). By contrast, NHEJ directly joins the two ends of a DSB, regardless of the sequence template at the exposed ends of the break, making it error-prone but available at all times during the cell cycle. NHEJ is involved in the maturation of immune cells through V(D)J recombination and class-switch recombination (24). The major steps in DSB-mediated repair pathways will be discussed here.

3.1. DNA damage-induced chromatin relaxation

In most eukaryotic cells the DNA is tightly packaged into the DNA-protein complex known as chromatin, which represents a significant barrier for DSB-repair proteins to access and repair DNA breaks. The dynamic restructuring of chromatin surrounding the lesion including

modifications of histone tails and remodelling of chromatin by remodelling factors allow HR and NHEJ machinery access to the damaged DNA [for more details, see (25-27)]. The most prominent chromatin modification after DSB induction is phosphorylation of H2AX (a type of histone H2A variant), which plays a primary role in the DNA damage repair by facilitating the access of HR factors to sites of DNA damage (discussed in the next section).

In response to a DSB, the chromatin surrounding the DSB is rapidly PARylated (modified by covalent addition of poly-ADP ribose, or PAR), a reaction catalysed by PARP1 (poly [ADP-ribose] polymerase 1) (28). This creates PAR chains at DSBs, allowing rapid and transient accumulation of the NuRD, PcG (polycomb group) and ALC1 remodelling complexes through interaction with PAR (28-30), and of the KAP-1/HP1 complex possibly through interaction with PAR at break sites [reviewed in (25)]. The NuRD complex is required for subsequent steps in DDR such as efficient marking of DNA damage site with ubiquitin by RNF8 (ring finger protein 8) and RNF168 (ring finger protein 168), and also for recruitment of BRCA1 to damaged DNA (31). PcG proteins exist in the form of two main complexes, PRC1 and PRC2 (polycomb repressive complex 1 and 2), which are recruited to DSB sites in a PARP-dependent manner. PRC1 can monoubiquitinate histone H2A at sites of DSBs, and PRC1-mediated monoubiquitination is required for subsequent RNF8- and/or RNF168-mediated polyubiquitination at DSBs (32-34). ALC1 may have a role in repositioning DSB-flanking nucleosomes, and in stabilizing the chromatin structure for further DSB processing and repair, while KAP-1/HP1 may promote the unpacking of heterochromatin, thereby facilitating repair of heterochromatic DSBs (26). These three complexes are retained at DSBs for only a short period of time, and then rapidly released from the chromatin, potentially through dePARylation by PARG (polyADP-ribose glycohydrolases) (25). The requirement of PARG for efficient DNA repair suggests that the presence of PAR at sites of DNA of damage must be tightly regulated.

Subsequent DSB signalling, including ATM activation and phosphorylation of histone H2AX, recruits MDC1 (mediator of DNA damage checkpoint 1) which then interacts with and loads another chromatin-remodelling complex, NuA4, onto chromatin adjacent to DSBs (35). Loading of NuA4 catalyses the exchange of H2A for H2A.Z through the p400 component of NuA4 [(36); reviewed in (25)]. This reaction is required for the acetylation of histone H4 by the TIP60 (also known as KAT5) component of NuA4, leading to the relaxation of DNA in proximity to DSBs [reviewed in (25, 37)].

Chromatin relaxation in both HR and NHEJ also involves ubiquitination of histone H2B by the heterodimer consisting of RNF20 (ring finger protein 20) and RNF40 (ring finger protein 40) in an ATM-dependent manner (38). Two tumour suppressors, CDC73 (cell division cycle 73) (39) and Smurf2 (Smad ubiquitin regulatory factor 2) (40), have been reported to regulate this ubiquitination

reaction, and this may represent a major mechanism by which mutations in these tumour suppressors exert their tumorigenic effect (39, 40).

3.2. Homologous recombination (HR)

HR occurs through a series of steps involving DSB-induced chromatin relaxation; recruitment of early HR factors to site of DSBs; DSB end resection; formation of the D loop; processing of the D loop or Holliday junctions; and the single-strand annealing (SSA) sub-pathway. We consider these in order.

3.2.1. Recruitment of early HR factors to DSBs

HR-mediated repair begins with the recognition and binding of DSB ends by the MRN (MRE11-RAD50-NBS1) complex (41, 42) (Figure 1*a*). Subsequently, MRN recruits a complex of ATM and the histone acetyltransferase TIP60 [TIP60/NuA4 complex mentioned above, which binds to histone H3 methylated at Lys-9 (H3K9me3)], to the sites of damage (43, 44). Both its recruitment to DSBs and phosphorylation of TIP60 by c-Abl kinase (20) are required to trigger the acetyltransferase activity of TIP60, leading to the activation of ATM by acetylation-induced auto-phosphorylation (44). The activated ATM then phosphorylates a multitude of substrates in response to DNA damage, particularly H2AX (termed γH2AX when phosphorylated), which serves as an anchoring platform for the accumulation of subsequent HR factors (Figure 1*b*), and is considered as an early marker of DSB signalling [reviewed in (45, 46)]. The recruitment of HR factors at sites of damage is regulated by various post-translational modifications which have been subject of comprehensive reviews (47, 48), only some of the most relevant post-translational modifications will be highlighted here.

The adaptor protein MDC1 localises to DSB sites by direct binding to γH2AX [reviewed in (9, 45)]. MDC1 also harbours a binding site for NBS1 component of MRN complex, promoting additional ATM recruitment and kinase activation (49, 50). The ability of MDC1 to bind γH2AX and NBS1 simultaneously enables positive feed-forward phosphorylation of H2AX by ATM and generates a megabase-sized γH2AX region surrounding DSBs [reviewed in (9, 46)].

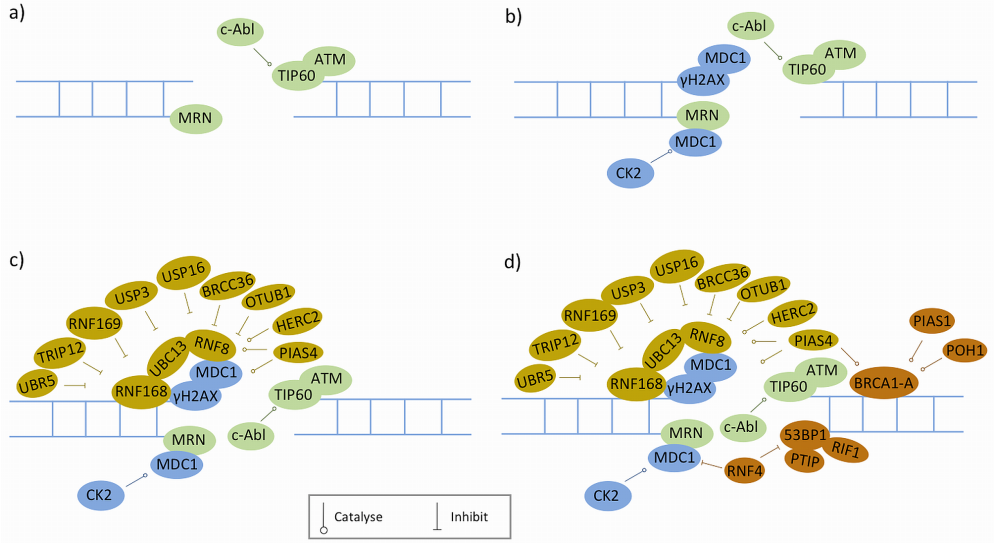


Figure 1 Recruitment of early homologous recombination (HR) factors to double-strand breaks (DSBs). Proteins represented in different colours are recruited at different times *a*) The MRN (MRE11-RAD50-NBS1) complex recognises and binds to DSBs, which then recruits ATM and TIP60. *b*) Activated ATM phosphorates H2AX, leading to the formation of γH2AX that provides binding sites for MDC1. *c*) Next, two ubiquitin ligases RNF8 and RNF168 are recruited to catalyse poly-ubiquitination of γH2AX. This ubiquitination event is tightly controlled by various positive and negative regulators. *d*) Subsequently, BRCA1 (in the form of BRCA1-A complex) and 53BP1 are recruited; these two proteins play important roles in the balance between HR and NHEJ, wherein a variety of regulatory mechanisms are involved.

After its recruitment, MDC1 is phosphorylated by ATM. MDC1 serves an important role as a scaffold for the downstream recruitment of the ubiquitin (Ub) E3 ligases RNF8 and RNF168, which work in tandem to ubiquitylate histone H2A and possibly other factors to create docking sites for Ub-binding proteins [for recent reviews, see (51, 52)]. Among these are 53BP1 (p53-binding protein 1) and Rap80/Abraxas, whose crucial function is to recruit BRCA1 (breast cancer type 1 susceptibility protein) to DSBs (53). Both of these two proteins are tumour suppressors and play a critical role in the pathway choice between HR and NHEJ (discussed in more detail below). The mechanisms for signal amplification exist due to crosstalk within one pathway and also across different pathways. RNF168 itself has ubiquitin-binding domain and E3 ligase activity, which together provide RNF168 the capability to amplify its own catalytic product. RNF8 but not RNF168 also promotes extensive decondensation of higher-order chromatin structure by recruiting the NuRD component CHD4 (31), which in turn promotes the recruitment and activation of RNF8, RNF168 and subsequent assembly of downstream repair factors [reviewed in (54)]. As discussed in section 3.1, PARylation is also required to recruit NuRD to assist chromatin ubiquitination at sites of breaks.

Multiple regulators tightly control RNF8/RNF168-mediated ubiquitination in HR. At present, four DUB enzymes (USP3, USP16, BRCC36 and OTUB1) and two HECT E3 ligases (TRIP12 and UBR5) have been shown to target RNF168 for proteasome-mediated degradation, potentially constraining the DSB repair machinery around the break site, and terminating the signal after repair has finished [reviewed in (51, 52)]. Interestingly, unlike TRIP12 and UBR5,

another HECT E3-ligase, HERC2, promotes RNF8/RNF168-based ubiquitination (55). In addition, another E3 ligase, RNF169, has an unexpected negative role in regulating RNF8/RNF168-induced ubiquitin signalling by directly binding to ubiquitin-modified chromatin, leading to impaired recruitment of 53BP1 and BRCA1 (52). Moreover, SUMOylation of HERC 2 and RNF8 is also involved in the regulation of RNF8/RNF168-induced ubiquitination (56).

Following RNF8/RNF168-catalysed ubiquitination of DSB-flanking chromatin, BRCA1 and 53BP1, two seemingly antagonistic factors, localise to the DSBs at approximately the same time (Figure 1*d*), providing an important layer of discrimination for DSB repair pathway choice. BRCA1 is required for functional HR, while 53BP1 promotes NHEJ by preventing DSB-end resection that is essential for HR. Interestingly, loss of 53BP1 can largely relieve the requirement of BRCA1 for HR, suggesting that a major role of BRCA1 in HR is to overcome a barrier to resection posed by 53BP1 (57, 58). This finding may have clinical implications, as a recent study showed that loss of BRCA1 often activates 53BP1 degradation in *BRCA1*-deficient cancer cells (59). Below we summarise current knowledge on how these two proteins are recruited, their role in determining pathway choice, and the regulation mechanisms that are involved.

BRCA1 participates in multiple stages of HR by forming at least three mutually exclusive complexes: the BRCA1-A, BRCC and BRCA1-C complexes by binding of different adaptors (Abraxas, BACH1 and CtIP, respectively) [reviewed in (60, 61)]. Following RNF8/RNF168-mediated ubiquitination of H2A and H2AX, BRCA1 is recruited to DSBs in the form of the BRCA1-A complex (61, 62). The accumulation of this complex to DSBs takes place through the binding of the Abraxas-RAP80 sub-complex with K63 poly-ubiquitin chains catalysed by RNF8 and RNF168 (63-65). SUMOylation of BRCA1 mediated by PIAS1 and PIAS4 is thought to promote the recruitment of the BRCA1-A complex, and stimulates the ubiquitin ligase activity of BRCA1 (66, 67).

53BP1 does not contain any known ubiquitin-binding motif and its accumulation at DSBs relies on binding to methylated histone H4 (68) and ubiquitinated histone H2A, the latter being a product of the RNF168 ubiquitin ligase activity (69). In addition, post-translational modifications of p53BP1 itself, including PIAS1/PIAS4-mediated SUMOylation, can promote the recruitment of 53BP1 at sites of DSBs (67)

The regulation of DSB repair pathway choice comes from the actions of 53BP1 and RIF1. Several recent studies have elegantly demonstrated that RIF1 is a downstream effector of 53BP1 in this process. In G1, RIF1 is recruited to DSB sites via ATM-dependent 53BP1 phosphorylation, and the 53BP1-RIF1 pathway inhibits the recruitment of BRCA1 to damage sites via an unknown mechanism to ensure repair through NHEJ. However, in S and G2 phases, CDK-and ATM-dependent phosphorylations of CtIP (CtBP-interacting protein) support the

formation of the CtIP-MRN-BRCA1 (BRCA1-C) complex which displaces RIF1 at break sites to promote DNA resection (70-73). However, unlike 53BP1, the loss of RIF1 only partially rescues HR defect in *BRCA1*-deficient cells, suggesting that additional RIF1-independent activities of 53BP1 might exist. Accordingly, a recent study (74) showed that PTIP is required for 53BP1-mediated inhibition of HR in *BRCA1*-deficient cells, but is dispensable for NHEJ during CSR (class switch recombination). Thus RIF1 and PTIP separate 53BP1 functions in productive and pathological DSB repair (74).

Compared to the mechanisms that regulate the assembly of early HR repair factors at DSB sites, those that regulate their disassembly remain largely unknown. The mechanism best-documented so far is the removal of MDC1 from DSB sites through PIAS4-mediated SUMOylation and consequent ubiquitination by the SUMO-targeted E3 ubiquitin ligase RNF4 (75-77), which leads to MDC1 degradation. MDC1 removal is important to remove 53BP1 from the damage sites, and is required for the recruitment of downstream HR proteins such as CtIP, RPA (replication protein A) and RAD51 (DNA repair protein RAD51 homolog 1) (75-77). In addition, TIP60-dependent histone H4 acetylation, which reduces the binding of 53BP1 to methylated histone H4 leads to reduced 53BP1 association with DSB-flanking chromatin (78).

3.2.2. DSB end resection

The sequential recruitment of early-stage HR factors, as outlined above, is required for and followed by DSB end resection – an evolutionarily conserved process that involves 5'-to-3' nucleolytic degradation of DSB ends to generate 3′ overhangs (a long stretch of single-stranded DNA (ssDNA) at DSB ends; also known as the 3' tail) (Figure 2). This 3′ overhang is a key determinant of DSB repair pathway choice, which commits cells to HR and is also required for activation of the ATR-mediated checkpoint response (79).

A two-step model has been put forward to describe DSB end resection in mammals (80). The first step, initiation of resection, involves a limited resection that removes ~50-100 nucleotides from the DSB ends, creating a short 3′ overhang that is further processed in the second step of resection generating a long 3′ overhang that is essential for the strand invasion step in HR [discussed below; for a recent review see (81)].

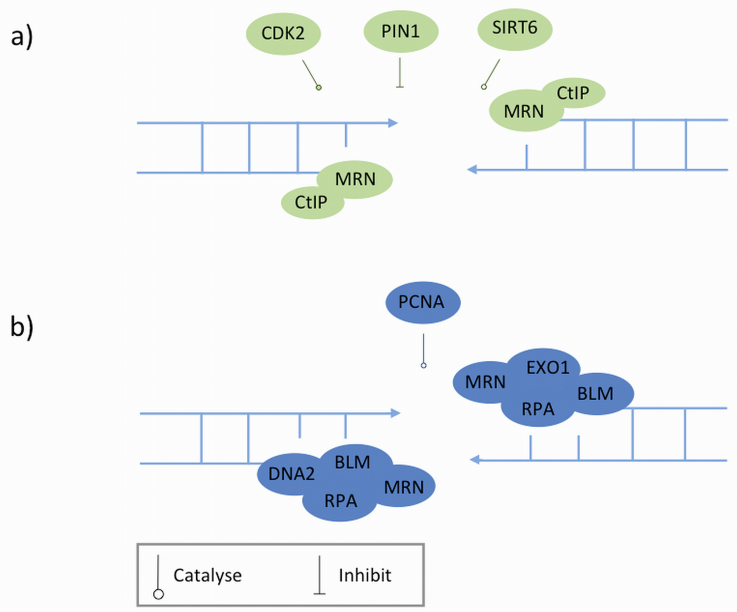


Figure 2 A two-step model for the Double Strand Break (DSB) end resection. Proteins represented in different colours are recruited at different stages. *a*) The first step, "initial resection", is carried out by the endonuclease activity of the MRN (MRE11-RAD50-NBS1) complex and promoted by CtIP. Multiple regulatory mechanisms, especially the cell cycle-dependent regulation are involved. *b*) The second step, "long-range resection", is performed by EXO1, or BLM in concert with DNA2.It remains unclear whether EXO1 and BLM work in parallel or interact.

The major resection machinery involved in first step is the MRN complex, which has an essential role in damage detection and ATM signalling, in conjunction with CtIP (82, 83) (Figure 2*a*). The initial resection *per se* is carried out by the endonuclease activity of the MRN complex followed by its exonuclease activity (84). CtIP promotes initial resection by interacting with MRN (79) and stimulating its endonuclease activity (83). The activity of CtIP in HR is regulated by multiple mechanisms, among which cell cycle-dependent regulation is of greatest importance because DSB resection must be restricted to the S and G2 phases where sister chromatids are present to serve as templates for HR. In the G1 phase, the level of CtIP protein is suppressed by proteasome-mediated degradation, which is subsequently alleviated as cells enter S phase (85). During S and G2 phases, CtIP is phosphorylated by CDKs (cyclin-dependent kinases) on multiple sites that promote resection in distinct ways. Among them, serine 327 is required for the CtIP-BRCA1 interaction and the formation of the BRCA1-C complex (82, 86), and threonine 847 for the localization of CtIP to DSBs and for end resection (87). These CDK-mediated phosphorylation signals directly link the DNA resection capacity with cell cycle control, thereby ensuring that the operation of HR is restricted to the S and G2 phases.

A phosphorylation-specific prolyl-isomerase, PIN1 (peptidyl-prolyl cis-trans isomerase NIMA-interacting 1), has recently been shown to counteract CDK-dependent end resection (88). PIN1 controls CtIP levels by promoting its isomerization in a CDK2-dependent manner followed by poly-ubiquitination (through an as-yet-unknown E3 ubiquitin ligase) and consequent degradation to limit end resection (88).

The second step, long-range resection, is carried out by two alternative pathways involving either the exonuclease function of EXO1 (DNA exonuclease I) alone, or the helicase function of BLM (Bloom syndrome, RecQ helicase-like) in concert with the nuclease function of DNA2 (DNA replication helicase 2) (89-91) (Figure 2*b*). It remains controversial whether BLM and EXO1 pathway work in parallel (90) or interact [(89); reviewed in (81)]. Recently, CDK1/2 has been shown to promote long-range resection by directly phosphorylating EXO1 on 4 different sites in mammalian cells (92).

Although PCNA (proliferating cell nuclear antigen) has recently been proposed to facilitate long-range resection by promoting the function of EXO1 (93), in general the regulatory mechanisms involved in this step are not well-understood. Interestingly, PCNA is also involved in base excision repair (BER) (94), nucleotide excision repair (NER) (95), mismatch repair (96), translesion synthesis (97), the Fanconi anaemia (FA) pathway (98) and the DNA repair synthesis step as well as suppressing inappropriate recombination in HR (99) (discussed below).

3.2.3. D loop formation and DNA repair synthesis

The 3′ overhang formed by end resection is coated and stabilised by RPA, which prevents ssDNA from forming secondary structure, and then RPA is displaced by the evolutionarily conserved recombinase RAD51. The loading of RAD51 onto ssDNA is a critical step in HR, as it generates a nucleoprotein filament that searches for and invades a nearby homologous duplex DNA template (usually a sister chromatid). As a consequence of this invasion, the second strand of the sister chromatid is displaced and a transient structure known as the D (displacement) loop is formed [reviewed in (100, 101)] (Figure 3*a*).

The loading of RAD51 onto ssDNA is promoted and controlled by multiple mechanisms [for a recent review, see (101)]. BRCA2 is the major recombinase accessory factor (also known as recombination mediator) that facilitates the loading of RAD51 onto ssDNA by overcoming the inhibitory effect of RPA (102). PALB2 is a partner and localiser of BRCA2, and serves as a molecular adaptor between BRCA1 and BRCA2 (103, 104). In this complex, BRCA1 is thought to fine-tune HR in part through its modulatory role in the PALB2-dependent loading of the BRCA2-RAD51 repair machinery at DNA breaks (103, 104). In addition, DSS1 (deleted in split hand/split foot 1), which forms a complex with BRCA2, is required for the stability of BRCA2 and facilitates the role of BRCA2 in RAD51–ssDNA filament formation (105, 106).

Recently, the SWI5-MEI5 complex was identified as an evolutionarily conserved mediator of RAD51 (107). This complex contributes to maintenance of the RAD51 nucleofilament in its active ATP-bound form by promoting the release of ADP from this structure (108).

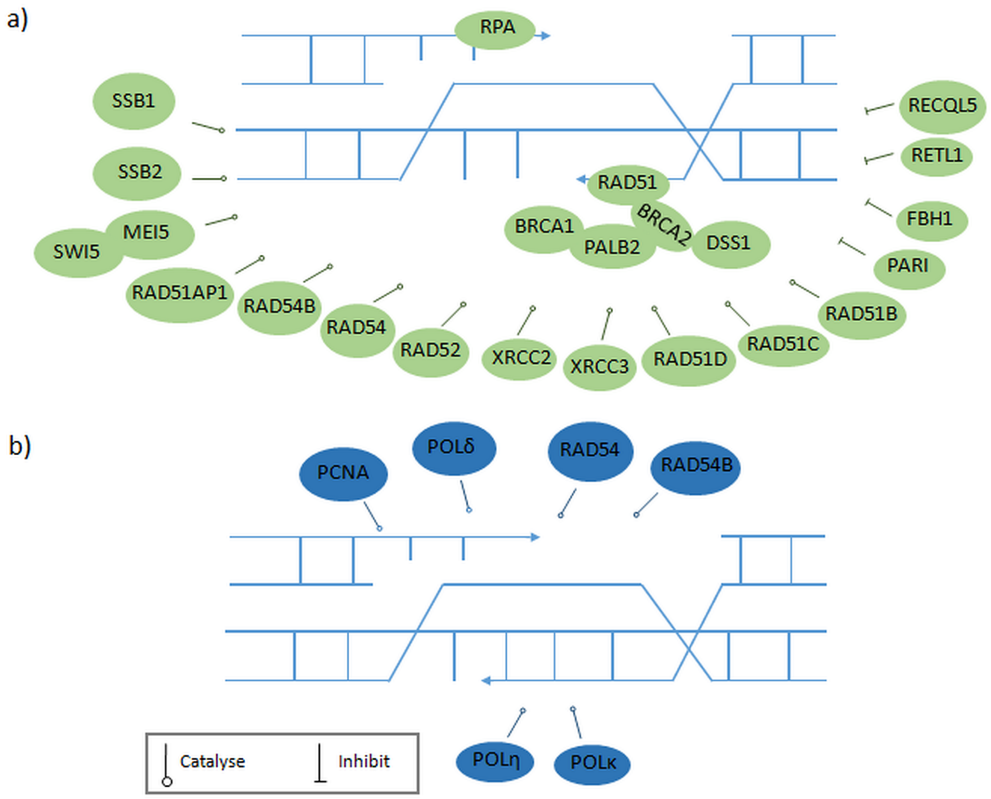


Figure 3 D loop formation and DNA repair synthesis. Proteins represented in different colours are recruited at different times *a*) The 3′ ssDNA overhang generated by DSB end resection is coated and stabilised by RPA, which is then displaced by RAD51 with the help of recombination mediators which promote both the formation and stability of RAD51-ssDNA filament. The balancing act of proteins involved in stability and dismantling of RAD51 filaments is depicted here as discussed in the text. Rad51 presynaptic filament performs homology searches with help of other proteins and invades nearby homologous duplex DNA template, resulting in the formation of the D loop structure. *b*) The invading strand is then elongated by copying missing genetic information from the template molecule, which involves the participation of several redundant DNA polymerases.

The loading of RAD51 onto ssDNA and subsequent formation of the D loop also depends on the concerted action of other proteins, which include the five RAD51 paralogs (RAD51B, RAD51C, RAD51D, XRCC2 and XRCC3) [reviewed in (109)], RAD52 [RAD52 homolog (*S. cerevisiae*)] (110), RAD54 [RAD54 homolog (*S. cerevisiae*)] and its paralog RAD54B [RAD54 homolog B (*S. cerevisiae*)] [reviewed in (111)], RAD51AP1 (RAD51 associated protein 1) (112, 113), and the two ssDNA-binding proteins SSB1 (single-strand DNA-binding protein 1) and SSB2 (single-strand DNA-binding protein 2) (114, 115).

Although HR has a key role in maintaining genome stability, its inappropriate activity can cause genomic instability potentially even leading to cancer. Several anti-recombinases suppress uncontrolled HR activity. These include PARI (PCNA-associated recombination inhibitor), RTEL1 (regulator of telomere elongation helicase 1), RECQL5 (RecQ protein-like 5) and FBH1 (F-box DNA helicase 1). PARI can disrupt toxic RAD51-ssDNA filaments in a PCNA-dependent manner (116), and overexpressed PARI has been implicated in the development of pancreatic cancer (117). RECQL5 regulates HR by targeting undesirable RAD51-ssDNA filament, and is important for tumour suppression in mice (118). FBH1 also functions by targeting RAD51-ssDNA filaments, and

its activity in HR is tightly controlled by PCNA (119, 120). RTEL1 can suppress inappropriate HR by promoting D loop disassembly (121).

Following D loop formation, the 3′ end of the invading strand serves as a primer for elongation of this strand *via* copying missing genetic information from the template molecule (100) (Figure 3*b*). For elongation to start, RAD51 in the 3′ end of the invading strand must be removed by RAD54 and RAD54B to reveal the 3′ hydroxyl group for priming (111). The DNA replication machinery involved in this elongation has not been well characterised. Recently Sebesta *et al.* showed that replicative DNA polymerase δ and two TLS polymerases (η and κ) play redundant roles in strand extension, and PCNA may act as a regulatory point for the recruitment of various polymerases and recombination outcomes (99).

HR can take two alternative routes beyond this point (Figure 4). Most frequently, in mitotic cells, elongation of the invading strand continues over only a limited distance, it is then released and anneals with the complementary ssDNA strand associated with the other DSB end. DSB repair is subsequently completed by gap-filling DNA synthesis and ligation. This sub-pathway is referred to as the SDSA (synthesis-dependent strand annealing) pathway [for a review, see (100)]. RTEL1 is the major enzyme that promotes the release of the invading strand by promoting the disassembly of the D loop structure, resulting in non-crossover products (no exchange of genetic information between the original DNA molecule and the template DNA molecule) (121). The D loop can also be processed by BLM to generate a non-crossover product (122), or by the MUS81-EME1 complex to generate a crossover product (123, 124)

Alternatively, in the DSBR (DSB repair) sub-pathway typical of meiosis, the second end of the DSB is captured to form an intermediate that harbours two Holliday junctions (HJs) [for reviews, see (7, 100)]. Processing/resolution of the HJ is promoted by various redundant enzymes including the BLM-TOPOIII-RMI1-RMI2 complex (125) and the endonucleases GEN1 (GEN endonuclease 1) (126), the MUS81-EME1 complex (123, 124) and the SLX1-SLX4 complex (127) (SLX4 is also known as FANCP in FA). In mitotic cells, the BLM-TOPOIII-RMI1-RMI2 complex is the major machinery responsible for dissolution of HJs to generate a non-crossover product (128, 129). Alternatively, HJs can be resolved by endonucleases that simply cleave HJs to generate crossover or non-crossover products. A recent study suggests two redundant pathways of HJ resolution in human cells, one pathway involves GEN1 and the other involves the coordinated action of SLX1-SLX4 and MUS81-EME1 (130). However, another recent study indicated that GEN1 alone cannot replace the resolvase activity provided by SLX1-SLX4 and MUS81-EME1 (131).

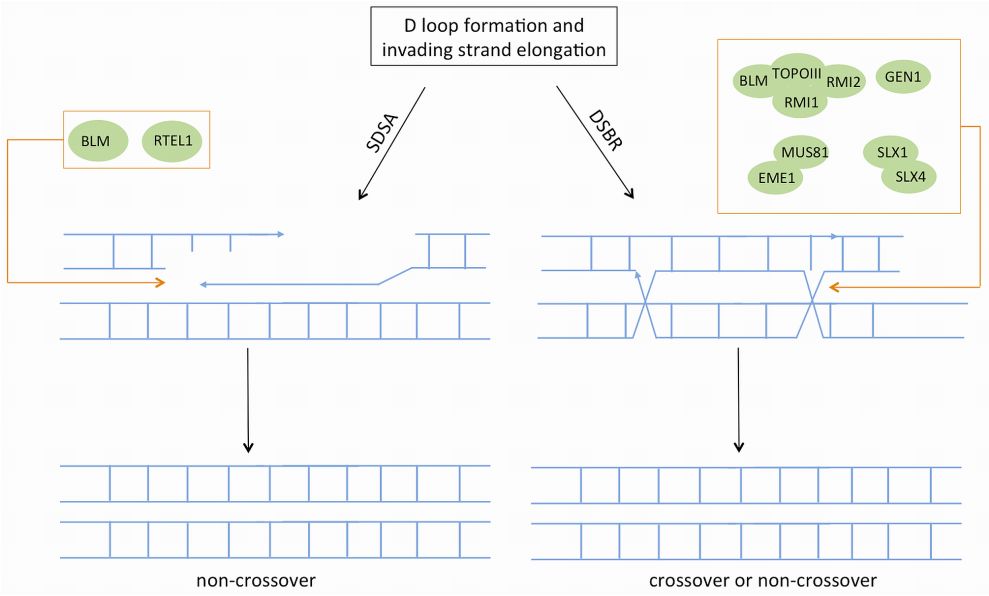


Figure 4 The SDSA (synthesis-dependent strand annealing) and DSBR (DSB repair) sub-pathways. D loop formation and DNA repair synthesis can follow two different routes namely SDSA and DSBR to complete homologous recombination. In SDSA invading strand is displaced from D-loop and annealed with complementary strand associated with second end of the DSB. SDSA is preferred over DSBR during mitosis, and mainly results in a non-crossover product. In the DSBR pathway, the other end of the DSB is captured and double Holliday Junction (dHJ) intermediate is formed which is then resolved to produce cross-over (mainly during meiosis) or non-crossover products.

3.2.4. The SSA sub-pathway

In addition to canonical HR, an alternative error-prone form of HR called SSA has been described (Figure 5). SSA is efficient in repairing DSBs between two direct repeat sequences flanking the ends of the DSB, and results in deletion of sequence between the two repeats. This pathway can be important for both DNA repair and mutagenesis, given that almost half of the human genome consists of repeated sequences (7, 100). The activity of SSA has been observed to increase in *BRCA2* or *RAD51*-deficient cells (132).

SSA is initiated by RAD52 that binds the 3' ssDNA ends generated by DSB end resection (the same process as described in Section 3.2.2), and functions in concert with RPA to facilitate strand annealing between the two direct repeats (133). This is followed by the removal of non-homologous 3′ single-stranded flaps between the two repeats (Figure 5), which is catalysed by a XPF-ERCC1 heterodimer that harbours 5′-3′ structure-specific endonuclease activity (134). In addition to SSA, XPF-ERCC1 also plays an important role in other DNA repair pathways including NER, FA and A-NHEJ (alternative-NHEJ; discussed below). The final step of SSA is the ligation of the two DSB ends, which is carried out by LIG3 (DNA ligase III) (135).

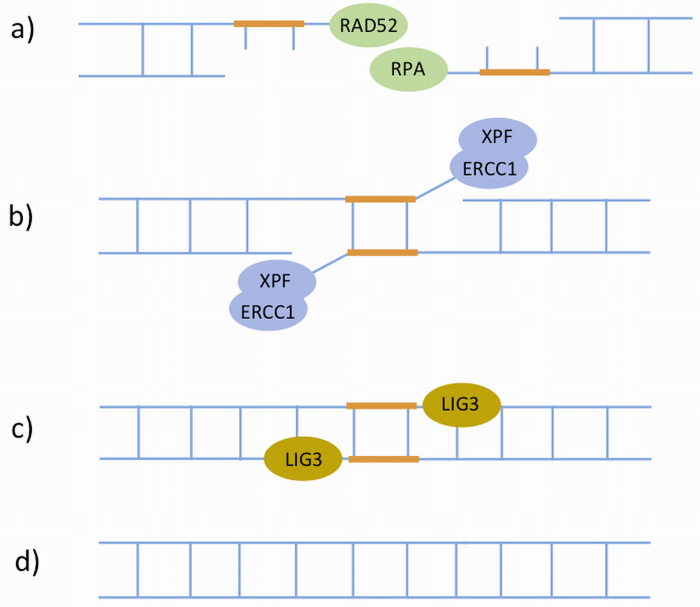


Figure 5 The Single-Strand Annealing (SSA) sub-pathway of Homologous Recombination. This is a Rad51-independent sub-pathway of HR, which operates when there are regions of homology/direct repeats at both sides of the DSB, allowing annealing. *a*) SSA is initiated by RAD52 that binds the 3' ssDNA ends generated by DSB end resection. RAD52 then functions in concert with RPA to facilitate strand annealing between the two direct repeats. *b*) Next, the XPF-ERCC1 heterodimers remove the non-homologous 3′ single-stranded flaps between the two repeats. *c*) The two DSB ends are re-joined by DNA ligase III. *d*) The sequence continuity is restored.

3.3. Non-homologous end-joining

NHEJ repairs the majority of DSBs throughout the cell cycle in human cells, although it remains unclear why such a low-fidelity pathway has evolved to dominate DSB repair. It is now generally accepted that there exist two forms of NHEJ: canonical NHEJ (C-NHEJ) and A-NHEJ.

3.3.1. Canonical NHEJ (C-NHEJ)

The most common amongst the two pathways, C-NHEJ (136-140) (Figure 6) commences with the rapid recognition and binding of the Ku heterodimer (consisting of Ku70 and Ku80) to DSBs (139, 141), which protects and stabilises the DNA ends, and serves as a scaffold onto which other NHEJ factors can dock (139).

Once Ku is bound to DSB ends, it directly recruits the DNA-PKcs kinase (DNA-dependent protein kinase catalytic subunit) to the damage sites (142), leading to activation of the kinase activity of DNA-PKcs (138, 139, 143). It has been shown *in vitro* that DNA-PKcs can phosphorylate a large number of NHEJ proteins, but *in vivo* only Artemis (144) and DNA-PKcs itself (auto-phosphorylation) (142) have been demonstrated so far as true substrates of DNA-PKcs phosphorylation [reviewed in (138)].

Ku also directly recruits a complex composed of XRCC4 (X-ray cross complementing protein 4), DNA ligase IV and XLF (XRCC4-like factor) (145, 146) to ligate DNA ends. This recruitment is independent of the presence of DNA-PKcs (146). XRCC4 has no known enzymatic

activity in NHEJ, and may serve as a second scaffold for the recruitment of other DSB-processing enzymes in this pathway. In addition, XRCC4 and XLF can form a filament that may play a role in bridging DSB ends (138, 147).

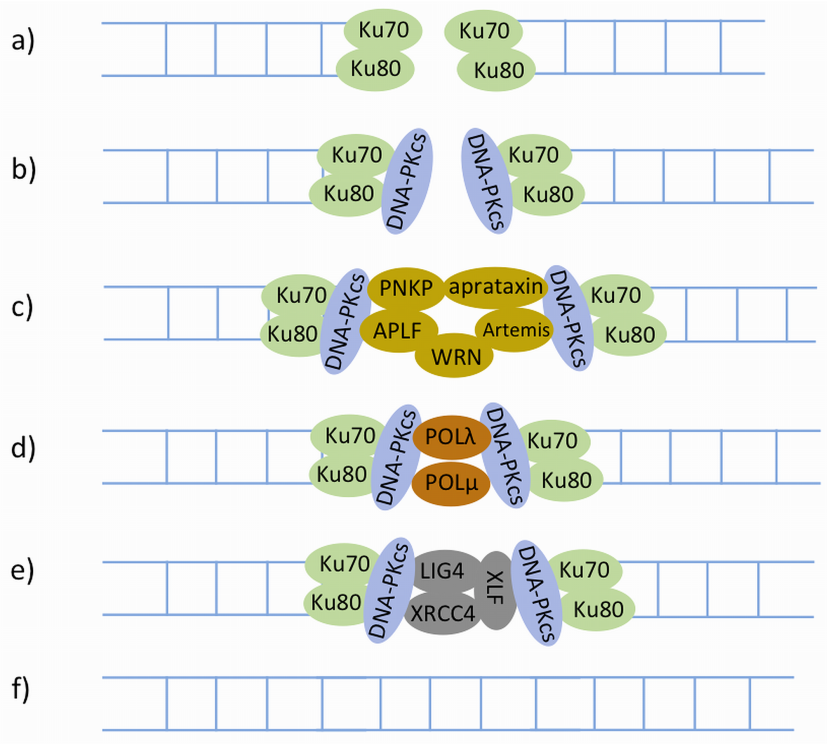


Figure 6 The Canonical NHEJ (C-NHEJ). Proteins represented in different colours are recruited at different stages *a*) The C-NHEJ pathway is initiated by the Ku70-Ku80 heterodimer. *b*) The Ku70-80 dimer then recruits the DNA-PKcs kinase. *c*) In many instances ends of the breaks are not amenable to direct ligation and must be resected or filled in prior to ligation by end processing factors depicted here are discussed in the text *d*) The synthesis step is catalysed by DNA polymerase μ and λ. *e*) The gap after DNA repair synthesis is ligated by the XRCC4-LIG4-XLF complex. *f*) The sequence continuity is restored.

In many instances the ends of a DSB are not amenable to direct ligation. For instance, the 5′ hydroxyls or 3′ phosphate termini of a DSB may be covalently modified or the ends may harbour 5’ or 3’ overhangs that must be resected or filled in prior to ligation. Important end-processing factors include PNKP (polynucleotide kinase-phosphatase), aprataxin, Ku, APLF (aprataxin-and-PNK-like factor), Artemis, WRN (Werner syndrome), and DNA polymerases μ and λ [reviewed in (138)]. Specifically PNKP (148), aprataxin (149) and Ku (150) remove blocking end groups such as non-ligatable 5′ hydroxyls or 3′ phosphates, as well as abasic sites near DSBs. APLF (151), Artemis (152, 153) and WRN (154) have roles in resecting DNA ends [reviewed in (138)]. APLF also facilitates the recruitment and/or retention of the XRCC4-DNA ligase IV-XLF complex at DSBs (155).

Following the removal of blocking end groups and DNA end resection, the resulting DNA gaps are filled by the action of DNA polymerase μ and λ, and are then ligated by LIG4 (DNA ligase IV) in conjunction with XRCC4 and XLF to finalise this pathway (156).

3.3.2. Alternative NHEJ (A-NHEJ)

Like C-NHEJ, A-NHEJ (Figure 7) has no inherent mechanism to ensure the restoration of the original DNA sequence in the vicinity of DSBs. Initial evidence for the existence of an alternative form of C-NHEJ, termed A-NHEJ, emerged when C-NHEJ is disabled [for reviews see (136, 137, 140, 157)], but recent studies have shown that substantial activity of this pathway can be observed when HR and C-NHEJ are still functional (158). A-NHEJ often benefits from microhomology in the proximity of DSBs; it has been frequently referred to as microhomology-mediated end-joining (MMEJ), but not all A-NHEJ requires microhomology for function (159).

A-NHEJ is initiated by PARP1, which competes with Ku for binding to DSB ends (160, 161). Following this binding, MRN, CtIP and BRCA1 are recruited to the damage sites for end resection (162-166), but this process can be blocked by 53BP1 to promote C-NHEJ to increase repair accuracy (167, 168). The step that finalises A-NHEJ is ligation. Unlike C-NHEJ, which exclusively utilises LIG4, ligation in A-NHEJ can be carried out by either LIG3 (169, 170) in a complex with XRCC1 (171), or LIG1 (DNA ligase I) (170, 172).

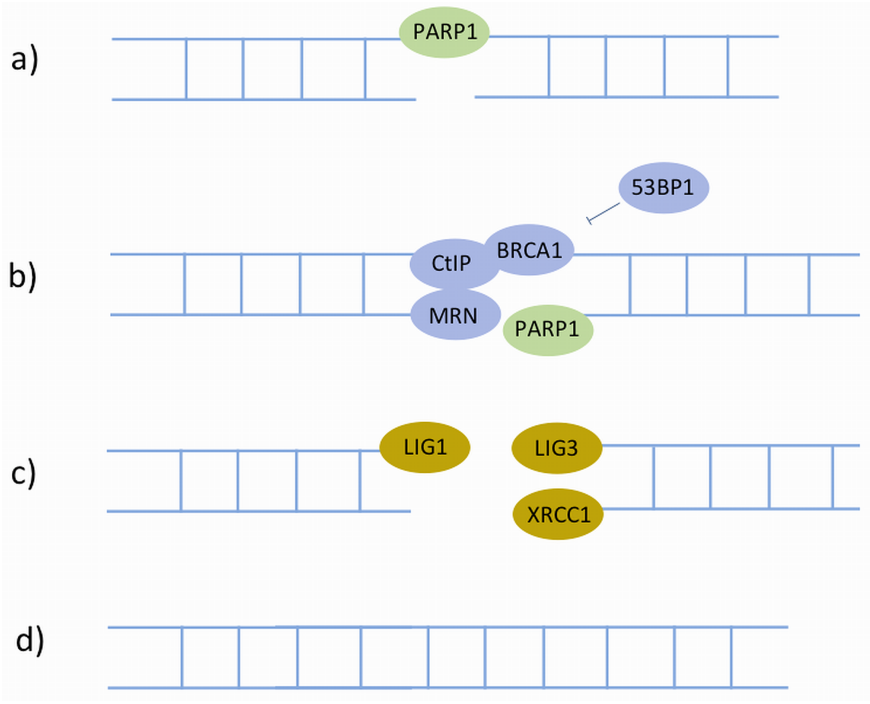


Figure 7 The Alternative A-NHEJ (A-NHEJ). Proteins represented in different colours are recruited at different stages. In A-NHEJ, *a*) the broken ends are detected and bound by PARP1. *b*) This is followed by end-processing by MRN, CtIP and BRCA1, which is prohibited by 53BP1. *c*) The ligation step can be performed by either LIG3 in concert with XRCC1, or LIG1. *d*) The sequence continuity is restored.

3.4. DSB-repair proteins in replication fork restart

A major physiological source of DNA damage in all cells and at every cell cycle is DNA replication. Replication forks are vulnerable to stalling or collapse (disassembly) due to obstacles encountered during replication, which can be unrepaired DNA damage or presence of DNA-bound proteins or secondary structures. A stalled fork is capable of resuming replication (replication fork restart), whereas a collapsed fork has become inactivated, possibly converting into DSBs that are repaired by HR. While a complex set of pathways from the core replication as well as fork-restart machinery are involved in the resumption of replication, several members of DSB-repair pathways, in particular HR, are known to be involved in this process to varying extents. The roles of these proteins here are distinct from the conventional HR activated during the S-phase. A detailed description of replication fork restart is beyond the scope of this article, and readers are directed to excellent reviews (23, 173, 174); here we summarise the roles of DSB-repair proteins in this process.

In case of shorter stalls (2 – 4 hours), most replication forks resume progression, with restart promoted by the proteins BLM, WRN, SMARCAL1, PARP1, XRCC3 and RAD51 (23). However, replication forks stalled for many hours (24 hours or more) are collapsed and DSBs are generated by the MUS81-EME1 complex (175), following which replication is resumed by new origin firing. The DSBs so-formed promote RAD51-dependent SDSA repair. In addition, PARP1, MRE11, BLM and WRN promote restart of collapsed forks. This suggests that DSB formation by MUS81, and DSB repair-mediated fork restart might be a mechanism to achieve replication fork progression, especially after prolonged fork stalling.

4. Implications of DNA repair for tumorigenesis and cancer therapy

At its core, cancer is a disease driven by genomic instability, accumulating into aberrations in large regions of the genome. Many of these aberrations are hallmarks of erroneous joining of DSB ends, resulting from disruption of DNA repair machineries. These defects, in turn acquired through certain ‘driver’ events such as mutations, copy-number changes or chromosomal rearrangements, that cause inactivation of DNA-repair, tumour-suppressor and apoptotic genes, leading to deficiency, misrepair or defects in the repair of DNA damage. Therefore an in-depth characterisation of the DSB-repair mechanisms (Section 3) and associating DSB-repair genes to specific driver events in cancer is crucial to understand cancer mechanisms and develop novel therapeutic strategies.

4.1. The genomic landscape of breast cancer

Germline mutations in DNA repair genes are major contributors to familial breast and ovarian cancer development (Table 1). For example, recent estimates suggest that 55-65% of women who inherit a deleterious *BRCA1* mutation, and around 45% who inherit a deleterious *BRCA2* mutation, will develop breast cancer by the age of 70 (176, 177). Patients who carry *BRCA1*/*2* mutations are also at a higher risk of developing contralateral disease (178). Likewise, germline mutations in *ATM* result in the autosomal recessive disorder Ataxia-telangiectasia, a neurodegenerative disorder characterised by hypersensitivity to ionizing radiation and a 100-fold increased risk of developing cancer (179). Heterozygous carriers of certain mutations in *ATM* also have a moderate risk of developing breast cancer (180).

The initiating events in sporadic cancer are less-clearly understood, but large-scale integrated molecular profiling of cancer genomes is beginning to reveal complex landscapes of point mutations, copy-number alterations and chromosomal rearrangements that contribute to tumorigenesis (3, 4, 181-186).

4.1.1. Point mutations and copy-number alterations

At the time of writing, the latest census on cancer mutations from COSMIC (http://cancer.sanger.ac.uk/cancergenome/projects/census/) (187) shows 19 genes implicated in breast cancer either by germline or somatic mutations, of which 11 are involved in DDR (Table 1). This list will expand as potential driver genes identified from large-scale sequencing initiatives are validated. For example, The Cancer Genome Atlas (182) identified 35 significantly mutated genes in breast cancer from analysis of 507 tumour genomes, including ten novel genes *TBX3*, *RUNX1*, *CBFB*, *AFF2*, *PIK3R1*, *PTPN22*, *PTPRD*, *NF1*, *SF3B1* and *CCND3*. This cohort included genomes harbouring deleterious germline variants in breast cancer susceptibility genes involved in DDR (*ATM*, *BRCA1*, *BRCA2*, *CHEK2*, *PTEN*, and *TP53*) (Table 1). Similar large-scale sequencing efforts (4, 182-186) have demonstrated extreme heterogeneity in mutation profiles, with *TP53* and *PIK3CA* being the most frequently mutated genes, occurring in over 30% of breast tumours, and the remaining genes (*e.g. GATA3*, *CDH1*, *MAP3K1*, *MAP2K4*, *MLL3*, *PTEN*, *AKT1*, *CDKN2A* and *NCOR1*) mutated at frequencies of 10% or less.

Table 1: DSB-repair and/or cell-cycle checkpoint genes associated with breast cancer development, compiled from TCGA and COSMIC. Germline mutations or epigenetic changes associated with breast cancer risk have been observed for some of these genes, while a few also fall close to single-nucleotide polymorphisms (SNPs) linked to breast cancer risk, identified from genome-wide association studies (GWAS) (http://www.genome.gov/gwastudies (242).

Gene	Gene name	Function of encoded protein	Chromosome band	Somatic mutation frequency in TCGA (%)	Somatic mutation frequency in COSMIC (%)	Copy-number alterations frequency in TCGA (%)	Target of germline mutations, epigenetic changes or SNPs (GWAS locus)
TP53	Tumour protein p53	Tumour suppressor involved in cell cycle arrest, apoptosis, senescence and DNA repair	17p13.1	23.15	29.0	0.60↓	Germline (243, 244)
MLL3	Myeloid/lymphoid or mixed-lineage leukaemia 3	Part of ASCOM complex regulated by acetylation toinduce expression of p53 targets such as p21 in response to DDR (245, 246)	7q36.1	4.61	6.48	0.40↑	
BRCA2	Familial breast/ovarian cancer gene 2	HR-mediated DSB repair	13q12.3	2.79	2.81	1.70↑↓	Germline (247) and GWAS locus
PTEN	Phosphatase and tensin homolog	Tumour suppressor with role in DNA repair through interactions with Chk1 and P53 pathways and regulation of RAD51 activity	10q23.3	2.30	9.13	1.80↓	Germline (248)
ATM	Ataxia-Telangiectasia Mutated	Master controller of cellular responses to DNA damage, regulates various tumour suppressors including P53 and BRCA1	11q22-q23	2.06	6.18	0.70↑↓	Germline (180, 249); epigenetic silencing (250, 251)
BRCA1	Familial breast/ovarian cancer gene 1	Tumour suppressor with key roles in HR-mediated DSB repair	17q21	1.82	2.19	1.10↓	Germline (252); epigenetic silencing (253)
AKT1	v-akt murine thymoma viral oncogene homolog 1	Regulates components of apoptotic machinery, also checkpoint pathway through phosphorylation of CHK1 (241)	14q32.32	1.45	1.17	1.00↑	
RB1	Retinoblastoma gene	Tumour suppressor, mediates cell cycle arrest	13q14.2	1.21	4.64	1.30↓	Germline (254)
BRIP1	BRCA1 interacting protein C-terminal helicase 1	Involved in HR-dependent DNA repair by association with BRCA1	17q22.2	0.97	1.39	7.50↑	Germline (255) – not confirmed
CDKN1-B	Cyclin-dependent kinase inhibitor 1B	Cell-cycle progression at G1	12p13.1-p12	0.61	0.48	0.70↑	
CCND3	Cyclin D3	Regulates cell cycle G1/S transition	6p21.1	0.61	0.42	1.10↑	
HIST1H-2BC	Histone cluster 1, H2bc	Core histone playing roles in DNA repair, replication and chromosomal stability	6p22.1	0.48	0.42	1.00↑	
CHEK2	CHK2 checkpoint homolog (*S. pombe*)	Cell cycle arrest in response to DNA damage. Interacts and phosphorylates BRCA1for activating DNA repair	22q12.1	0.48	2.57	0.50↑	Germline (256)
EP300	300 kDa E1A-Binding protein gene	Regulates transcription *via* chromatin remodelling. Regulated by acetylation in response to DDR (257)	22q13.2	0.36	2.98	0	
BAP1	BRCA1 associated protein-1 (ubiquitin carboxy-terminal hydrolase)	Binds to BRCA1 and involved in cell cycle growth, response to DNA damage and chromatin dynamics.	3p21.1	0.24	2.97	0.40↓	Germline (258) – not confirmed
CCND1	Cyclin D1	Regulates cell cycle during G1/S, also interacts with a network of repair proteins including RAD51 to regulate HR (259)	11q13	0.12	0.59	14.1↑	GWAS locus (260)
PALB2	Partner and localizer of BRCA2	Critical role in HR-mediated repair by recruiting RAD51 and BRCA2 to DSB sites.	16p12.2	0	1.14	1.80↑	Germline (261)

GWAS locus – if the gene is noted as the nearest gene to a breast cancer associated SNP identified by a GWAS study. However, it should be noted that unless a reference is given there is no evidence that the gene is the target of that association. Copy number alterations are shown as predominant amplification (↑) and homozygous deletion (↓) in TCGA cases.

In addition to point mutations, most solid tumours display widespread changes in chromosome number (aneuploidy), as well as deletions, inversions, translocations, and other genetic abnormalities. By integrated analysis of DNA copy-number alterations and gene expression profiles in 2000 breast cancers, Curtis *et al.* (183) identified 45 regions of the genome that act as copy-number drivers of gene expression in breast cancer. These included known (*MYC*, *CCND1*, *MDM2*, *ERBB2*, *CCNE1*) and putative candidate driver genes (*MDM1*, *MDM4*, *CDK3*, *CDK4*, *CAMK1D*, *PI4KB*, *NCOR1*, *PPP2R2A*, *MTAP* and *MAP2K4*).

4.1.2. Chromosomal rearrangements

Chromosomal rearrangements, particularly intra- and inter-chromosomal translocations, may fuse two genes to create an oncogene (*e.g. BCR-ABL* fusion gene in chronic myeloid leukaemia) or, in a small number of cases, inactivate a tumour suppressor gene (*e.g.TEL-AML* fusion repressing the tumour suppressor *TEL1*). Catastrophic rearrangements (*chromothripsis*), which affects local chromosomal regions, can also have similar tumorigenic effects (188-190). Chromothripsis is characterized by highly focal shattering of chromosomes into tens to hundreds of segments (188), leading to focal amplifications, deletions or fusions in chromosomal regions (191).

In an analysis of 24 breast tumours, rearrangements were found in known cancer genes including *BRAF*, *PAX3*, *PAX5*, *NSD1*, *PBX1*, *MSI2* and *ETV6*, each of which is a partner in a fusion gene in several other human cancers. Rearrangements were also found in tumour suppressor genes such as *RB*, *ABC* and *FBXW7*, possibly resulting in gene inactivation (3).

The analyses of rearrangements also revealed striking signatures of defective DNA repair by different pathways. For instance, in the same study of 1821 rearrangement junctions (3642 breakpoints) in 24 breast tumours (3), the segments on either side of each rearrangement junction showed overlapping microhomology immediately adjacent to the junction. Approximately 15% of the rearrangements showed non-templated sequence at the junction. Overlapping microhomology and non-templated sequences at rearrangement junctions are often considered to be signatures of the NHEJ-mediated repair process. In particular, in some of the tumour genomes, rearrangements with zero base pairs of microhomology were most frequent, while in others rearrangements with two or more base pairs were common, indicating at least two variants of NHEJ repair to be operative in different breast tumours. *BRCA1*- and *BRCA2*-associated tumours showed few tandem duplications, indicating that the mechanisms responsible for chromosomal rearrangements in these tumours were distinct from those in triple-negative tumours, which exhibited tandem duplications.

On the other hand, the mechanistic origin of chromothripsis is largely unclear. Although large-scale genome analyses have not identified chromothriptic rearrangements in breast tumours (192), analysis of rearranged regions in glioblastomas, bone and lung tumours have identified a

catastrophic event in which chromosomes undergo multiple fragmentation and rejoining, mainly by NHEJ (191, 193). Sequencing of samples from primary, relapse and metastatic tumours have noted that most of these chromothriptic events were present in the primary and initial tumours and did not necessarily occur in an on-going basis or only during metastasis (188, 194).

4.1.3. Molecular basis of breast tumours revealed through mutational signatures

Large-scale sequencing studies such as TCGA and those initiated by the International Cancer Genome Consortium (ICGC) have generated an increasingly comprehensive atlas of molecular alterations across a wide range of cancers and allowing a systematic exploration of the genetic basis of cancer. This has led to studies identifying *mutational signatures* across cancers (195, 196). For example, 21 mutational signatures have been identified across ~7000 tumours (195) associating cancers to risk factors such as exposure to specific carcinogens, particularly smoking in lung cancer and UV radiation in melanoma.

Breast tumours are largely characterised by three signatures (1B, 2 and 3) strongly associated with age, APOBEC activity and *BRCA1*/*2* mutations, respectively. These signatures are predominantly characterised by C>G and C>T changes, and "rainfall plot" clustering of these mutations exhibits heavily mutated stretches of the genome characterised by distinctive C>T transitions at TpCpX trinucleotides, resembling *kataegis* (Greek for shower or thunderstorm) in these plots (197, 198).

The correlation of breast cancer mutations with the age of diagnosis (Signature 1B) is consistent with the hypothesis that a substantial proportion of these mutations are acquired over the lifetime of the patient at a relatively constant rate that is similar in different people. Signature 2 is attributed to the overactivity the APOBEC family of cytidine deaminases, which convert cytidine to uracil, coupled to activity of the base excision repair and DNA replication machineries. Because APOBEC activation constitutes part of the innate immune response to viruses and retrotransposons, it has been hypothesised that collateral damage on the genome might be initiated from a response originally directed at retrotransposing DNA elements or exogenous viruses (199, 200). Finally, Signature 3 is associated with inactivating mutations in *BRCA1* and *BRCA2* genes, indicating that abrogation of functional HR- and/or NHEJ-mediated repair contributes considerably to breast cancer development, even in patients not harbouring a germline mutation in either of these two genes.

Likewise, another large-scale study (196) characterised ~3000 tumours on the basis of ~500 selected functional events (SFE) encompassing copy-number gains and losses, recurrent mutations and epigenetic silencing of genes. Based on these SFEs, tumours were classified into two classes, *M*

primarily with mutations, and *C* primarily with copy-number alterations, revealing a characteristic trend of "genome hyperbola" – cancers have either a large number of mutations or a large number of copy-number alternations, but rarely both. Breast cancer was included in class *C,* as reflected in amplifications of the *MYC* oncogene, *CCND1* and *PIK3CA*, deletion of *CDK2NA*, and inactivating mutations in *TP53* leading to copy-number instability. A subclass of tumours in *C* showed copy-number alterations in cell cycle regulation and DDR pathways attributable to amplification of the gene encoding the mitotic regulator AURKA kinase and the inactivation of *BRCA1* and *BRCA2* genes.

Analyses of mutational signatures across cancers have led to three fundamental observations so far (182, 195, 196): tumours originating in the same organ or tissue vary substantially in the number, type and pattern of genomic alterations; similar patterns of genomic alteration are observed in tumours from different tissues of origin; and common mutational signatures in tumours are "imprints" of common underlying mechanisms (*e.g.* APOBEC activity or DDR deficiency) or factors (*e.g.* age and exposure to carcinogens/DNA damage). These observations suggest that 'signature-driven therapies' designed and tailored to tissue-specific tumour types could be extensible across classes of cancer that share similar mutational signatures.

4.2. DNA repair pathways as targets for cancer therapy

The efficacy of DNA damage-based therapy can be modulated selectively towards cancer cells by targeting DNA-damage induced checkpoint and repair pathways (21, 201, 202). Drugs and agents that inhibit the activity of DNA-repair pathways have been reviewed in detail elsewhere (6, 21, 203, 204); here we focus on an exciting strategy called *synthetic lethality*, which has recently gained attention due to its potential for being both selective for and highly effective against cancer cells.

4.2.1. Synthetic lethality-based therapy

Synthetic lethality refers to a type of genetic interaction in which the co-occurrence of two genetic events results in death of the cell or organism (205, 206). For example, two genes are synthetic lethal when their simultaneous inactivation results in cell death, but deletion of either individually does not affect cell viability. Two common models have been proposed to explain synthetic lethality between two genes (207):

(i) the two genes function in parallel pathways, with each contributing to a process essential to viability, or

(ii) the genes encode proteins that form part of an essential complex that is partially functional in the absence of one of the proteins, but its functions are completely disrupted in the absence of both.

4.2.2. Leveraging synthetic lethality to selectively target cancer cells

Cancer cells undergo a multi-step selection for acquisition of hallmark phenotypes including evasion of apoptosis, insensitivity to growth-control signals and unlimited replicative potential (208, 209). In this scenario, genes of minor importance to the well-being of normal cells may become essential lifelines specifically in cancer cells, providing opportunities for novel therapeutic interventions (209).

The DNA repair machinery is attractive in this context, given that cancer cells are driven by a loss of fidelity in DNA repair and continually accumulate further DNA damage (Figure 8). Selective killing of cancer cells could be made possible either by targeting an otherwise non-essential gene that has turned essential and hence lethal specifically in cancer cells, or alternately by inducing massive amounts of DNA damage (*via* DNA-damaging chemotherapeutic agents or radiation) and subsequently forcing cancer cells into DNA-damage-induced apoptosis. Normal cells remain adequately buffered to repair the induced DNA damage, and will continue to maintain regular function and homeostasis.

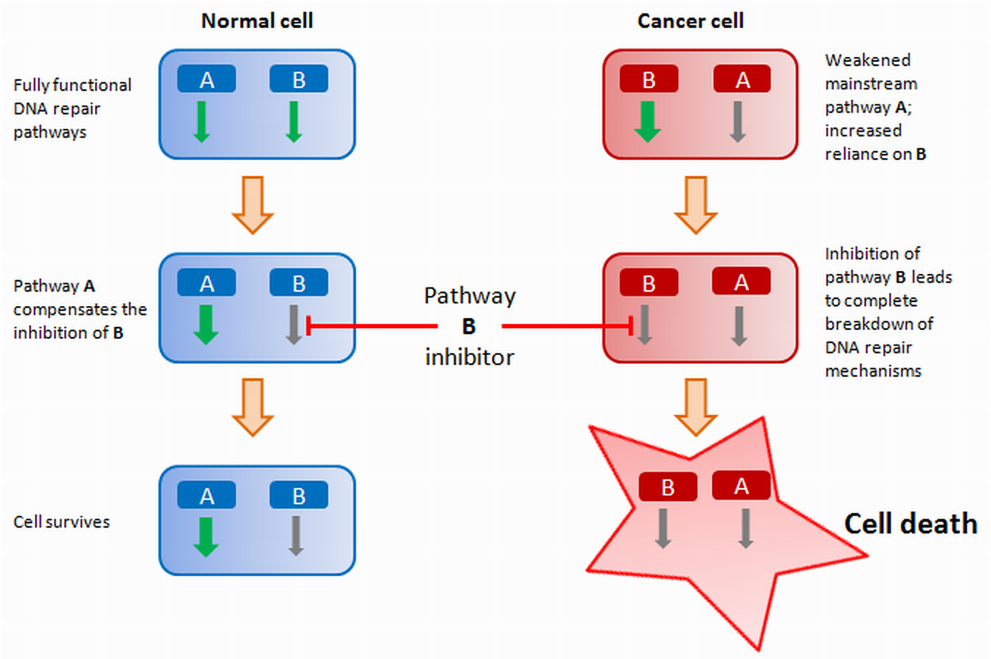


Figure 8 Strategy for synthetic lethality based cancer therapy. Targeted inhibition of DNA-damage repair pathways in defined cancer cell populations to selectively kill cancer cells.

4.2.3. BRCA1-PARP1 synthetic lethality

A clinically relevant synthetically lethal relationship in the DDR has been documented between mutations in *BRCA1* or *BRCA2* and the inhibition of PARPs (210, 211). *BRCA1*- or *BRCA2*-deficient cells are sensitive to siRNA-mediated knockdown or chemical inhibition of PARP, leading to the clinical testing of PARP inhibitors as potential anti-cancer drugs in *BRCA1* or *BRCA2*-deficient cancers. This suggests a new approach to cancer therapeutics: olaparib (AZD2281), veliparib (ABT-888) and niraparib (MK-4827) are some of the PARP inhibitors that are in advanced clinical trials (212).

Despite the pronounced synthetic lethality observed between BRCA1/2 deficiency and PARP inhibition, the exact mechanism responsible for this observed phenomenon remains somewhat contentious. Nonetheless, the inhibition of PARP itself is not lethal for mammals, and $PARP1^{-/-}$ mice are viable and fertile, even though they manifest accelerated aging and exhibit a higher incidence of tumours compared to wild-type controls (213). The reason PARP1 is non-essential could be due to overlapping functions with other members of the PARP family, in particular PARP2 (214). However, most PARP inhibitors inhibit both PARP1 and PARP2 and the side-effects of this inhibition appear to be mild in both mice and humans (212), suggesting that the pronounced effect of PARP inhibition might be specific to HR-deficient cells.

An early model attributed the pronounced lethality between BRCA1/2 deficiency and PARP inhibition to the involvement of PARP1 in BER. In this model, PARP inhibition leads to persistent accumulation of SSBs, which convert to lethal DSBs during the S-phase; the inability to repair these DSBs in HR-deficient cancer cells result in the selective death of these cells. However, subsequent studies failed to demonstrate an increase in SSBs upon PARP inhibition in BRCA2-deficient cells (215), or reproduce synthetic lethality upon inhibition of XRCC1, an essential component of BER (216), suggesting that this may not be the mechanism of action of this synthetically lethal relationship.

Recent studies suggest that additional roles for PARP in DNA repair may be responsible for this observed synthetic lethality (215-221). The contribution of PARP1 to DSB repair, in particular through its involvement in alternative NHEJ (Section 3.3.2), has been suggested for its observed synthetic lethality with HR. A deficiency in HR could further result in lesions that require PARP1-dependent NHEJ for repair. However, PARP inhibition shifts this dependency onto the DNA-PKcs-dependent canonical NHEJ, thereby exposing HR-deficient cells to aberrant repair, resulting in increased genomic instability and apoptosis (216, 221) (Figure 9).

In addition to these roles, PARP1 also plays a role at stalled replication forks (Section 3.4), and *in vitro* studies in BRCA2-deficient cells suggest that PARP1 protects stalled replication forks

from MRE11A-mediated degradation in a manner that is distinct and complementary to the role of BRCA2, resulting in synthetic lethality with BRCA2 at stalled replication forks (219, 220).

Further, the chemical action of PARP inhibitors itself can contribute to cell death. Most PARP inhibitors target the catalytic site of the enzyme and thereby block the binding to its substrates, thus preventing PAR-synthesis and causing the enzyme to be "trapped" on the DNA (222). As a result, PARP inhibition not only restricts its signalling, but the inactivated enzyme forms an obstacle that prevents access for repair proteins to the damaged site or hinders replication (223).

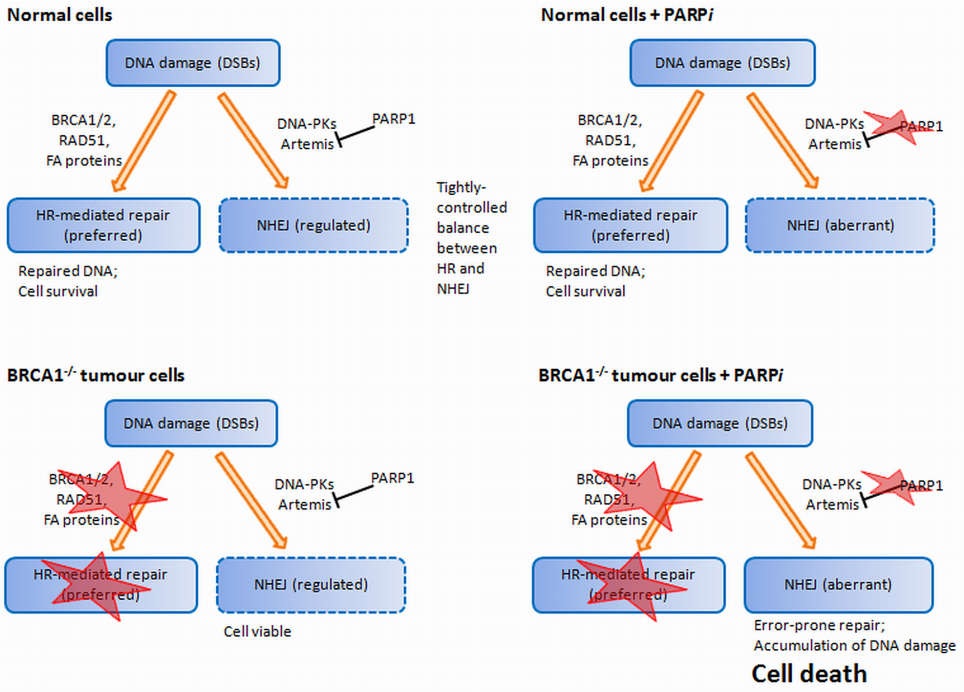


Figure 9 Alternative model (216) centred on the unrestricted error-prone NHEJ as the cause of death in tumour cells. HR-deficient cells were found to be hypersensitive to PARP1 inhibition, but this effect was reversed by disabling classical NHEJ, verified through knockdown of *Ku80* and *Artemis*. This suggests that classical NHEJ contributes to the toxicity of PARP1 inhibitors in HR-deficient cells, and therefore an active classical NHEJ is necessary for PARP inhibitor-based synthetic lethality.

In normal cells, the inhibition of PARP alone is not sufficient to kill these cells as both HR and the canonical NHEJ pathways provide functional repair of DSBs throughout the cell cycle. Cancer cells are prone to excessive oncogene-induced replication stress, often resulting in increased levels of DNA damage (224). An increased PARP activity might be required for protecting stalled replication forks from degradation, fork restart (Section 3.4) or alternative NHEJ-mediated repair of DSBs generated at replication fork, and the increased levels of PARP1 expression seen in cancer cells might be reflective of such PARP activity (210). Therefore, upon PARP inhibition, as demonstrated in BRCA1/BRCA2-deficient cells, HR becomes essential to resolve these lesions (211). Indeed cells lacking or with inhibited PARP1 display an increase in HR, sister chromatid exchange and micronuclei formation (225, 226). It is also possible that various components of HR are in general essential for survival during PARP inhibition, and thus become synthetically lethal to

the cell during HR deficiency. In support of this, deficiency in RAD51, MRE11, NBS1, RPA1 and loss of *PALB2* and *RAD51D* has been shown to sensitise cells to PARP inhibition (227).

4.3. DSB repair as a determinant of resistance to cancer therapy

It has long been known that DSB-repair-deficient tumours attain resistance by improving their DSB repair potential (5). In some cases such as breast and ovarian cancer, mutational events in any of the genes (Section 4.1) affect only a subset of the domains of these genes, leaving the remaining domains functional with some residual pathway activity. For example, mammary tumours from $BRCA1^{C61G}$ mutant mice lacking a functional RING domain respond more poorly to cisplatin than do BRCA1-*null* mammary tumours (228), indicating that a certain basal activity of RING-deficient BRCA1 protein is sufficient to reduce initial drug sensitivity and promote drug resistance (229).

Secondary mutations in these genes can potentially restore their functionality, also contributing to therapy resistance (230, 231). For example, *BRCA1*- and *BRCA2*-mutant cells are known to develop acquired resistance to PARP-inhibitor treatment due in part to secondary mutations in these genes that restore the reading frame and produce a functional protein that reverses the HR deficit (230-232). In some of the PARP-inhibitor resistant *BRCA2*-mutant clones the mutation was spliced out, allowing functional BRCA2 proteins to be produced with internal deletions (233, 234).

Tumours with intrinsic HR deficiencies may counteract therapeutic sensitivity by rewiring their DNA repair pathways or by altering pathway choices. For example, alterations in the balance between HR and NHEJ may change responses to DSB-inducing agents, as is seen when the loss of 53BP1 resulting from truncating *TP53BP1* mutations confers PARP-inhibitor resistance in *BRCA1*-deficient cells by providing the CtIP protein with unrestricted access to DNA breaks and facilitating DNA end resection (57, 58, 235). Loss of 53BP1 also restricts NHEJ, which is required for the success of PARP1-inhibitor therapy (216). Likewise, HSP90-mediated stabilisation of BRCT domain-mutated BRCA1 protein can confer resistant to PARP inhibitors, reversible by treatment with an HSP90 inhibitor (236). Suppressing NHEJ components including Ku70, Lig4 or DNA-PKcs alters the tight balance between HR and NHEJ, and such a strategy has the potential to be used against FA (237, 238).

These observations collectively mean that deeper understanding of the underlying functional relationships, particularly their specific genetic context and alternative rewiring in response to therapy, is critical to counter restoration of DSB repair and hence the development of resistance to therapy. Cancer pathways have been compared to a transport or subway map (209, 239): blocking a major commuter line will have repercussions throughout the network as passengers try to find

alternative routes to their destinations. Similarly, targeted cancer therapies are thwarted by the emergence of drug resistance, typically through unanticipated rewiring of signalling pathways and the surfacing of alternative functional relationships that are not obvious from the original wiring diagrams (209, 240, 241).

5. Conclusion

Aberrant DDR lies at the core of all cancers, and cancer cells differ genetically from normal cells in their ability to repair their DNA. These differences can be exploited to selectively kill cancer cells. However, this requires a deep understanding of the complexities of DDR pathways, in particular of DSB repair, in order to precisely modulate the pathways and sensitise cancer cells to DSB-inducing drugs.

Here we have presented an in-depth description of DSB repair mechanisms, focusing on HR and NHEJ, reflecting the latest state of knowledge in the field. We have discussed synthetic lethality as a new strategy to target components of these pathways, with emphasis on the BRCA1-PARP1 relationship that opened up promising avenues for targeted therapies in breast cancer. Finally we considered cases in which cancer cells become resistant to therapy by improving their DSB-repair potential. These observations suggest that we need better biomarkers to detect patients with HR deficiency eligible for treatment with PARP inhibitors. It is likely that the response to other cancer therapeutics including inhibitors of other repair pathways will also become more predictable, thus allowing more effective, targeted cancer treatments.

Funding

CL is supported by a University of Queensland International scholarship. GCT and KKK are supported by Senior Principal Research Fellowships from National Health and Medical Research Council (NHMRC). PTS is funded by a Fellowship from the National Breast Cancer Foundation, Australia. This work was funded by NHMRC Project Grant 1028742 to PTS and MAR.

References

1. Khanna, K.K. and Jackson, S.P. (2001) DNA double-strand breaks: signaling, repair and the cancer connection. *Nat. Genet.*, **27**, 247–254.

2. Rich, T., Allen, R.L. and Wyllie, A.H. (2000) Defying death after DNA damage. *Nature*, **407**, 777–783.

3. Stephens, P.J., McBride, D.J., Lin, M.-L., Varela, I., Pleasance, E.D., Simpson, J.T., Stebbings, L.A., Leroy, C., Edkins, S., Mudie, L.J., et al. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

4. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A. and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.

5. Fojo, T. (2001) Cancer, DNA repair mechanisms, and resistance to chemotherapy. *J. Natl. Cancer Inst.*, **93**, 1434–1436.

6. Lord, C.J. and Ashworth, A. (2012) The DNA damage response and cancer therapy. *Nature*, **481**, 287–294.

7. Ciccia, A. and Elledge, S.J. (2010) The DNA damage response: making it safe to play with knives. *Molecular Cell*, **40**, 179–204.

8. Surova, O. and Zhivotovsky, B. (2013) Various modes of cell death induced by DNA damage. *Oncogene*, **32**, 3789–3797.

9. Shiloh, Y. and Ziv, Y. (2013) The ATM protein kinase: regulating the cellular response to genotoxic stress, and more. *Nature Reviews Molecular Cell Biology*, **14**, 197–210.

10. Abraham, R.T. (2001) Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes & Development*, **15**, 2177–2196.

11. Matsuoka, S., Ballif, B.A., Smogorzewska, A., McDonald, E.R., Hurov, K.E., Luo, J., Bakalarski, C.E., Zhao, Z., Solimini, N., Lerenthal, Y., et al. (2007) ATM and ATR substrate analysis reveals extensive protein networks responsive to DNA damage. *Science*, **316**, 1160–1166.

12. Bensimon, A., Schmidt, A., Ziv, Y., Elkon, R., Wang, S.-Y., Chen, D.J., Aebersold, R. and Shiloh, Y. (2010) ATM-dependent and -independent dynamics of the nuclear phosphoproteome after DNA damage. *Sci Signal*, **3**, rs3.

13. Bennetzen, M.V., Larsen, D.H., Bunkenborg, J., Bartek, J., Lukas, J. and Andersen, J.S. (2010) Site-specific phosphorylation dynamics of the nuclear proteome during the DNA damage response. *Mol. Cell Proteomics*, **9**, 1314–1323.

14. Smith, J., Tho, L.M., Xu, N. and Gillespie, D.A. (2010) The ATM-Chk2 and ATR-Chk1 pathways in DNA damage signaling and cancer. *Adv. Cancer Res.*, **108**, 73–112.

15. Kuerbitz, S.J., Plunkett, B.S., Walsh, W.V. and Kastan, M.B. (1992) Wild-type p53 is a cell cycle checkpoint determinant following irradiation. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 7491–7495.

16. Macleod, K.F., Sherry, N., Hannon, G., Beach, D., Tokino, T., Kinzler, K., Vogelstein, B. and Jacks, T. (1995) p53-dependent and independent expression of p21 during cell growth, differentiation, and DNA damage. *Genes & Development*, **9**, 935–944.

17. Manke, I.A., Nguyen, A., Lim, D., Stewart, M.Q., Elia, A.E.H. and Yaffe, M.B. (2005) MAPKAP kinase-2 is a cell cycle checkpoint kinase that regulates the G2/M transition and S phase progression in response to UV irradiation. *Molecular Cell*, **17**, 37–48.

18. Warmerdam, D.O., Brinkman, E.K., Marteijn, J.A., Medema, R.H., Kanaar, R. and Smits, V.A.J. (2013) UV-induced G2 checkpoint depends on p38 MAPK and minimal activation of ATR-Chk1 pathway. *Journal of Cell Science*, **126**, 1923–1930.

19. Wang, X., Zeng, L., Wang, J., Chau, J.F.L., Lai, K.P., Jia, D., Poonepalli, A., Hande, M.P., Liu, H., He, G., et al.

(2011) A positive role for c-Abl in Atm and Atr activation in DNA damage response. *Cell Death and Differentiation*, **18**, 5–15.

20. Kaidi, A. and Jackson, S.P. (2013) KAT5 tyrosine phosphorylation couples chromatin sensing to ATM signalling. *Nature*, **498**, 70–74.

21. Curtin, N.J. (2012) DNA repair dysregulation from cancer driver to therapeutic target. *Nat Rev Cancer*, **12**, 801–817.

22. Liu, P., Carvalho, C.M.B., Hastings, P.J. and Lupski, J.R. (2012) Mechanisms for recurrent and complex human genomic rearrangements. *Curr. Opin. Genet. Dev.*, **22**, 211–220.

23. Petermann, E. and Helleday, T. (2010) Pathways of mammalian replication fork restart. *Nature Reviews Molecular Cell Biology*, **11**, 683–687.

24. Alt, F.W., Zhang, Y., Meng, F.-L., Guo, C. and Schwer, B. (2013) Mechanisms of programmed DNA lesions and genomic instability in the immune system. *Cell*, **152**, 417–429.

25. Price, B.D. and D'Andrea, A.D. (2013) Chromatin remodeling at DNA double-strand breaks. *Cell*, **152**, 1344–1354.

26. Goodarzi, A.A. and Jeggo, P.A. (2012) The heterochromatic barrier to DNA double strand break repair: how to get the entry visa. *Int J Mol Sci*, **13**, 11844–11860.

27. Bao, Y. (2011) Chromatin response to DNA double-strand break damage. *Epigenomics*, **3**, 307–321.

28. Chou, D.M., Adamson, B., Dephoure, N.E., Tan, X., Nottke, A.C., Hurov, K.E., Gygi, S.P., Colaiácovo, M.P. and Elledge, S.J. (2010) A chromatin localization screen reveals poly (ADP ribose)-regulated recruitment of the repressive polycomb and NuRD complexes to sites of DNA damage. *Proceedings of the National Academy of Sciences*, **107**, 18475–18480.

29. Ahel, D., Horejsí, Z., Wiechens, N., Polo, S.E., Garcia-Wilson, E., Ahel, I., Flynn, H., Skehel, M., West, S.C., Jackson, S.P., et al. (2009) Poly(ADP-ribose)-dependent regulation of DNA repair by the chromatin remodeling enzyme ALC1. *Science*, **325**, 1240–1243.

30. Gottschalk, A.J., Timinszky, G., Kong, S.E., Jin, J., Cai, Y., Swanson, S.K., Washburn, M.P., Florens, L., Ladurner, A.G., Conaway, J.W., et al. (2009) Poly(ADP-ribosyl)ation directs recruitment and activation of an ATP-dependent chromatin remodeler. *Proceedings of the National Academy of Sciences*, **106**, 13770–13774.

31. Luijsterburg, M.S., Acs, K., Ackermann, L., Wiegant, W.W., Bekker-Jensen, S., Larsen, D.H., Khanna, K.K., van Attikum, H., Mailand, N. and Dantuma, N.P. (2012) A new non-catalytic role for ubiquitin ligase RNF8 in unfolding higher-order chromatin structure. *EMBO J*, **31**, 2511–2527.

32. Ismail, I.H., Andrin, C., McDonald, D. and Hendzel, M.J. (2010) BMI1-mediated histone ubiquitylation promotes DNA double-strand break repair. *The Journal of Cell Biology*, **191**, 45–60.

33. Ginjala, V., Nacerddine, K., Kulkarni, A., Oza, J., Hill, S.J., Yao, M., Citterio, E., van Lohuizen, M. and Ganesan, S. (2011) BMI1 is recruited to DNA breaks and contributes to DNA damage-induced H2A ubiquitination and repair. *Molecular and Cellular Biology*, **31**, 1972–1982.

34. Ismail, I.H., McDonald, D., Strickfaden, H., Xu, Z. and Hendzel, M.J. (2013) A small molecule inhibitor of polycomb repressive complex 1 inhibits ubiquitin signaling at DNA double-strand breaks. *Journal of Biological Chemistry*, **288**, 26944–26954.

35. Xu, Y., Sun, Y., Jiang, X., Ayrapetov, M.K., Moskwa, P., Yang, S., Weinstock, D.M. and Price, B.D. (2010) The p400 ATPase regulates nucleosome stability and chromatin ubiquitination during DNA repair. *The Journal of Cell Biology*, **191**, 31–43.

36. Xu, Y., Ayrapetov, M.K., Xu, C., Gursoy-Yuzugullu, O., Hu, Y. and Price, B.D. (2012) Histone H2A.Z controls a critical chromatin remodeling step required for DNA double-strand break repair. *Molecular Cell*, **48**, 723–733.

37. Sun, Y., Jiang, X. and Price, B.D. (2010) Tip60: connecting chromatin to DNA damage signaling. *Cell Cycle*, **9**, 930–936.

38. Moyal, L., Lerenthal, Y., Gana-Weisz, M., Mass, G., So, S., Wang, S.-Y., Eppink, B., Chung, Y.M., Shalev, G., Shema, E., et al. (2011) Requirement of ATM-dependent monoubiquitylation of histone H2B for timely repair of DNA double-strand breaks. *Molecular Cell*, **41**, 529–542.

39. Hahn, M.A., Dickson, K.-A., Jackson, S., Clarkson, A., Gill, A.J. and Marsh, D.J. (2012) The tumor suppressor CDC73 interacts with the ring finger proteins RNF20 and RNF40 and is required for the maintenance of histone 2B monoubiquitination. *Human Molecular Genetics*, **21**, 559–568.

40. Blank, M., Tang, Y., Yamashita, M., Burkett, S.S., Cheng, S.Y. and Zhang, Y.E. (2012) A tumor suppressor function of Smurf2 associated with controlling chromatin landscape and genome stability through RNF20. *Nat Med*, **18**, 227–234.

41. Lamarche, B.J., Orazio, N.I. and Weitzman, M.D. (2010) The MRN complex in double-strand break repair and telomere maintenance. *FEBS Letters*, **584**, 3682–3695.

42. Stracker, T.H. and Petrini, J.H.J. (2011) The MRE11 complex: starting from the ends. *Nature Reviews Molecular Cell Biology*, **12**, 90–103.

43. Sun, Y., Jiang, X., Xu, Y., Ayrapetov, M.K., Moreau, L.A., Whetstine, J.R. and Price, B.D. (2009) Histone H3 methylation links DNA damage detection to activation of the tumour suppressor Tip60. *Nat. Cell Biol.*, **11**, 1376–1382.

44. Sun, Y., Jiang, X., Chen, S., Fernandes, N. and Price, B.D. (2005) A role for the Tip60 histone acetyltransferase in the acetylation and activation of ATM. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13182–13187.

45. Bhatti, S., Kozlov, S., Farooqi, A.A., Naqi, A., Lavin, M. and Khanna, K.K. (2011) ATM protein kinase: the linchpin of cellular defenses to stress. *Cell. Mol. Life Sci.*, **68**, 2977–3006.

46. van Attikum, H. and Gasser, S.M. (2009) Crosstalk between histone modifications during the DNA damage response. *Trends in Cell Biology*, **19**, 207–217.

47. Ulrich, H.D. (2012) Ubiquitin and SUMO in DNA repair at a glance. *Journal of Cell Science*, **125**, 249–254.

48. Bologna, S. and Ferrari, S. (2013) It takes two to tango: Ubiquitin and SUMO in the DNA damage response. *Front Genet*, **4**, 106.

49. Spycher, C., Miller, E.S., Townsend, K., Pavic, L., Morrice, N.A., Janscak, P., Stewart, G.S. and Stucki, M. (2008) Constitutive phosphorylation of MDC1 physically links the MRE11-RAD50-NBS1 complex to damaged chromatin. *The Journal of Cell Biology*, **181**, 227–240.

50. Chapman, J.R. and Jackson, S.P. (2008) Phospho-dependent interactions between NBS1 and MDC1 mediate chromatin retention of the MRN complex at sites of DNA damage. *EMBO reports*, **9**, 795–801.

51. Bartocci, C. and Denchi, E.L. (2013) Put a RING on it: regulation and inhibition of RNF8 and RNF168 RING finger E3 ligases at DNA damage sites. *Front Genet*, **4**, 128.

52. Tomimatsu, N., Mukherjee, B., Hardebeck, M.C., Ilcheva, M., Camacho, C.V., Harris, J.L., Porteus, M., Llorente, B., Khanna, K.K. and Burma, S (2014). Phosphorylation of EXO1 by CDKs 1 and 2 regulates DNA end resection and repair pathway choice. *Nature communications*, **5**, 3561 doi: 10.1038/ncomms4561.

53. Polo, S.E. and Jackson, S.P. (2011) Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications. *Genes & Development*, **25**, 409–433.

54. Altmeyer, M. and Lukas, J. (2013) To spread or not to spread--chromatin modifications in response to DNA damage. *Curr. Opin. Genet. Dev.*, **23**, 156–165.

55. Bekker-Jensen, S., Rendtlew Danielsen, J., Fugger, K., Gromova, I., Nerstedt, A., Lukas, C., Bartek, J., Lukas, J. and Mailand, N. (2010) HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes. *Nat. Cell Biol.*, **12**, 80–6– sup pp 1–12.

56. Danielsen, J.R., Povlsen, L.K., Villumsen, B.H., Streicher, W., Nilsson, J., Wikström, M., Bekker-Jensen, S. and Mailand, N. (2012) DNA damage-inducible SUMOylation of HERC2 promotes RNF8 binding via a novel SUMO-

binding Zinc finger. *The Journal of Cell Biology*, **197**, 179–187.

57. Bunting, S.F., Callen, E., Wong, N., Chen, H.-T., Polato, F., Gunn, A., Bothmer, A., Feldhahn, N., Fernandez-Capetillo, O., Cao, L., et al. (2010) 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. *Cell*, **141**, 243–254.

58. Bouwman, P., Aly, A., Escandell, J.M., Pieterse, M., Bartkova, J., van der Gulden, H., Hiddingh, S., Thanasoula, M., Kulkarni, A., Yang, Q., et al. (2010) 53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers. *Nature Structural & Molecular Biology*, **17**, 688–695.

59. Grotsky, D.A., Gonzalez-Suarez, I., Novell, A., Neumann, M.A., Yaddanapudi, S.C., Croke, M., Martinez-Alonso, M., Redwood, A.B., Ortega-Martinez, S., Feng, Z., et al. (2013) BRCA1 loss activates cathepsin L-mediated degradation of 53BP1 in breast cancer cells. *The Journal of Cell Biology*, **200**, 187–202.

60. Greenberg, R.A., Sobhian, B., Pathania, S., Cantor, S.B., Nakatani, Y. and Livingston, D.M. (2006) Multifactorial contributions to an acute DNA damage response by BRCA1/BARD1-containing complexes. *Genes & Development*, **20**, 34–46.

61. Huen, M.S.Y., Sy, S.M.H. and Chen, J. (2010) BRCA1 and its toolbox for the maintenance of genome integrity. *Nature Reviews Molecular Cell Biology*, **11**, 138–148.

62. Harris, J.L. and Khanna, K.K. (2011) BRCA1 A**-**complex fine tunes repair functions of BRCA1. *Aging (Albany NY)*, **3**, 461–463.

63. Sobhian, B., Shao, G., Lilli, D.R., Culhane, A.C., Moreau, L.A., Xia, B., Livingston, D.M. and Greenberg, R.A. (2007) RAP80 targets BRCA1 to specific ubiquitin structures at DNA damage sites. *Science*, **316**, 1198–1202.

64. Wang, B., Matsuoka, S., Ballif, B.A., Zhang, D., Smogorzewska, A., Gygi, S.P. and Elledge, S.J. (2007) Abraxas and RAP80 form a BRCA1 protein complex required for the DNA damage response. *Science*, **316**, 1194–1198.

65. Kim, H., Chen, J. and Yu, X. (2007) Ubiquitin-binding protein RAP80 mediates BRCA1-dependent DNA damage response. *Science*, **316**, 1202–1205.

66. Morris, J.R., Boutell, C., Keppler, M., Densham, R., Weekes, D., Alamshah, A., Butler, L., Galanty, Y., Pangon, L., Kiuchi, T., et al. (2009) The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress. *Nature*, **462**, 886–890.

67. Galanty, Y., Belotserkovskaya, R., Coates, J., Polo, S., Miller, K.M. and Jackson, S.P. (2009) Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks. *Nature*, **462**, 935–939.

68. Botuyan, M.V., Lee, J., Ward, I.M., Kim, J.-E., Thompson, J.R., Chen, J. and Mer, G. (2006) Structural basis for the methylation state-specific recognition of histone H4-K20 by 53BP1 and Crb2 in DNA repair. *Cell*, **127**, 1361–1373.

69. Fradet-Turcotte, A., Canny, M.D., Escribano-Diaz, C., Orthwein, A., Leung, C.C.Y., Huang, H., Landry, M.-C., Kitevski-LeBlanc, J., Noordermeer, S.M., Sicheri, F., et al. (2013) 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature*, **499**, 50–54.

70. Feng, L., Fong, K.-W., Wang, J., Wang, W. and Chen, J. (2013) RIF1 Counteracts BRCA1-mediated End Resection during DNA Repair. *Journal of Biological Chemistry*, **288**, 11135–11143.

71. Escribano-Diaz, C., Orthwein, A., Fradet-Turcotte, A., Xing, M., Young, J.T.F., Tkac, J., Cook, M.A., Rosebrock, A.P., Munro, M., Canny, M.D., et al. (2013) A cell cycle-dependent regulatory circuit composed of 53BP1-RIF1 and BRCA1-CtIP controls DNA repair pathway choice. *Molecular Cell*, **49**, 872–883.

72. Chapman, J.R., Barral, P., Vannier, J.-B., Borel, V., Steger, M., Tomas-Loba, A., Sartori, A.A., Adams, I.R., Batista, F.D. and Boulton, S.J. (2013) RIF1 is essential for 53BP1-dependent nonhomologous end joining and suppression of DNA double-strand break resection. *Molecular Cell*, **49**, 858–871.

73. Zimmermann, M., Lottersberger, F., Buonomo, S.B., Sfeir, A. and de Lange, T. (2013) 53BP1 regulates DSB repair using Rif1 to control 5' end resection. *Science*, **339**, 700–704.

74. Callen, E., Di Virgilio, M., Kruhlak, M.J., Nieto-Soler, M., Wong, N., Chen, H.-T., Faryabi, R.B., Polato, F., Santos, M., Starnes, L.M., et al. (2013) 53BP1 mediates productive and mutagenic DNA repair through distinct phosphoprotein interactions. *Cell*, **153**, 1266–1280.

75. Luo, K., Zhang, H., Wang, L., Yuan, J. and Lou, Z. (2012) Sumoylation of MDC1 is important for proper DNA damage response. *EMBO J*, **31**, 3008–3019.

76. Galanty, Y., Belotserkovskaya, R., Coates, J. and Jackson, S.P. (2012) RNF4, a SUMO-targeted ubiquitin E3 ligase, promotes DNA double-strand break repair. *Genes & Development*, **26**, 1179–1195.

77. Yin, Y., Seifert, A., Chua, J.S., Maure, J.-F., Golebiowski, F. and Hay, R.T. (2012) SUMO-targeted ubiquitin E3 ligase RNF4 is required for the response of human cells to DNA damage. *Genes & Development*, **26**, 1196–1208.

78. Tang, J., Cho, N.W., Cui, G., Manion, E.M., Shanbhag, N.M., Botuyan, M.V., Mer, G. and Greenberg, R.A. (2013) Acetylation limits 53BP1 association with damaged chromatin to promote homologous recombination. *Nature Structural & Molecular Biology*, **20**, 317–325.

79. Jazayeri, A., Falck, J., Lukas, C., Bartek, J., Smith, G.C.M., Lukas, J. and Jackson, S.P. (2006) ATM- and cell cycle-dependent regulation of ATR in response to DNA double-strand breaks. *Nat. Cell Biol.*, **8**, 37–45.

80. Symington, L.S. and Gautier, J. (2011) Double-strand break end resection and repair pathway choice. *Annu. Rev. Genet.*, **45**, 247–271.

81. Grabarz, A., Barascu, A., Guirouilh-Barbat, J. and Lopez, B.S. (2012) Initiation of DNA double strand break repair: signaling and single-stranded resection dictate the choice between homologous recombination, non-homologous end-joining and alternative end-joining. *Am J Cancer Res*, **2**, 249–268.

82. Chen, L., Nievera, C.J., Lee, A.Y.-L. and Wu, X. (2008) Cell cycle-dependent complex formation of BRCA1.CtIP.MRN is important for DNA double-strand break repair. *J. Biol. Chem.*, **283**, 7713–7720.

83. Sartori, A.A., Lukas, C., Coates, J., Mistrik, M., Fu, S., Bartek, J., Baer, R., Lukas, J. and Jackson, S.P. (2007) Human CtIP promotes DNA end resection. *Nature*, **450**, 509–514.

84. Shibata, A., Moiani, D., Arvai, A.S., Perry, J., Harding, S.M., Genois, M.-M., Maity, R., van Rossum-Fikkert, S., Kertokalio, A., Romoli, F., et al. (2014) DNA Double-Strand Break Repair Pathway Choice Is Directed by Distinct MRE11 Nuclease Activities. *Molecular Cell*, **53**, 7–18.

85. Germani, A., Prabel, A., Mourah, S., Podgorniak, M.-P., Di Carlo, A., Ehrlich, R., Gisselbrecht, S., Varin-Blank, N., Calvo, F. and Bruzzoni-Giovanelli, H. (2003) SIAH-1 interacts with CtIP and promotes its degradation by the proteasome pathway. *Oncogene*, **22**, 8845–8851.

86. Yu, X. and Chen, J. (2004) DNA damage-induced cell cycle checkpoint control requires CtIP, a phosphorylation-dependent binding partner of BRCA1 C-terminal domains. *Molecular and Cellular Biology*, **24**, 9478–9486.

87. Huertas, P. and Jackson, S.P. (2009) Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. *J. Biol. Chem.*, **284**, 9558–9565.

88. Steger, M., Murina, O., Hühn, D., Ferretti, L.P., Walser, R., Hänggi, K., Lafranchi, L., Neugebauer, C., Paliwal, S., Janscak, P., et al. (2013) Prolyl isomerase PIN1 regulates DNA double-strand break repair by counteracting DNA end resection. *Molecular Cell*, **50**, 333–343.

89. Nimonkar, A.V., Genschel, J., Kinoshita, E., Polaczek, P., Campbell, J.L., Wyman, C., Modrich, P. and Kowalczykowski, S.C. (2011) BLM-DNA2-RPA-MRN and EXO1-BLM-RPA-MRN constitute two DNA end resection machineries for human DNA break repair. *Genes & Development*, **25**, 350–362.

90. Tomimatsu, N., Mukherjee, B., Deland, K., Kurimasa, A., Bolderson, E., Khanna, K.K. and Burma, S. (2012) Exo1 plays a major role in DNA end resection in humans and influences double-strand break repair and damage signaling decisions. *DNA Repair*, **11**, 441–448.

91. Bolderson, E., Tomimatsu, N., Richard, D.J., Boucher, D., Kumar, R., Pandita, T.K., Burma, S. and Khanna, K.K. (2010) Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks. *Nucleic Acids Research*, **38**, 1821–1831.

92. Tomimatsu, N., Mukherjee, B., Hardebeck, M.C., Ilcheva, M., Camacho, C.V., Harris, J.L., Porteus, M., Llorente, B., Khanna, K.K. and Burma, S. Phosphorylation of EXO1 by CDKs 1 and 2 regulates DNA end resection and repair pathway choice. *Nat. Commun.*, **5**, 3561.

93. Chen, X., Paudyal, S.C., Chin, R.-I. and You, Z. (2013) PCNA promotes processive DNA end resection by Exo1. *Nucleic Acids Research*, 10.1093/nar/gkt672.

94. Matsumoto, Y. (2001) Molecular mechanism of PCNA-dependent base excision repair. *Prog. Nucleic Acid Res. Mol. Biol.*, **68**, 129–138.

95. Gilljam, K.M., Müller, R., Liabakk, N.B. and Otterlei, M. (2012) Nucleotide excision repair is associated with the replisome and its efficiency depends on a direct interaction between XPA and PCNA. *PLoS ONE*, **7**, e49199.

96. Umar, A., Buermeyer, A.B., Simon, J.A., Thomas, D.C., Clark, A.B., Liskay, R.M. and Kunkel, T.A. (1996) Requirement for PCNA in DNA mismatch repair at a step preceding DNA resynthesis. *Cell*, **87**, 65–73.

97. Kannouche, P.L., Wing, J. and Lehmann, A.R. (2004) Interaction of human DNA polymerase eta with monoubiquitinated PCNA: a possible mechanism for the polymerase switch in response to DNA damage. *Molecular Cell*, **14**, 491–500.

98. Geng, L., Huntoon, C.J. and Karnitz, L.M. (2010) RAD18-mediated ubiquitination of PCNA activates the Fanconi anemia DNA repair network. *The Journal of Cell Biology*, **191**, 249–257.

99. Sebesta, M., Burkovics, P., Juhasz, S., Zhang, S., Szabo, J.E., Lee, M.Y.W.T., Haracska, L. and Krejci, L. (2013) Role of PCNA and TLS polymerases in D-loop extension during homologous recombination in humans. *DNA Repair*, **12**, 691–698.

100. San Filippo, J., Sung, P. and Klein, H. (2008) Mechanism of eukaryotic homologous recombination. *Annu. Rev. Biochem.*, **77**, 229–257.

101. Krejci, L., Altmannova, V., Spirek, M. and Zhao, X. (2012) Homologous recombination and its regulation. *Nucleic Acids Research*, **40**, 5795–5818.

102. Holloman, W.K. (2011) Unraveling the mechanism of BRCA2 in homologous recombination. *Nature Structural & Molecular Biology*, **18**, 748–754.

103. Zhang, F., Fan, Q., Ren, K. and Andreassen, P.R. (2009) PALB2 functionally connects the breast cancer susceptibility proteins BRCA1 and BRCA2. *Molecular Cancer Research*, **7**, 1110–1118.

104. Sy, S.M.H., Huen, M.S.Y. and Chen, J. (2009) PALB2 is an integral component of the BRCA complex required for homologous recombination repair. *Proceedings of the National Academy of Sciences*, **106**, 7155–7160.

105. Li, J., Zou, C., Bai, Y., Wazer, D.E., Band, V. and Gao, Q. (2006) DSS1 is required for the stability of BRCA2. *Oncogene*, **25**, 1186–1194.

106. Liu, J., Doty, T., Gibson, B. and Heyer, W.-D. (2010) Human BRCA2 protein promotes RAD51 filament formation on RPA-covered single-stranded DNA. *Nature Structural & Molecular Biology*, **17**, 1260–1262.

107. Yuan, J. and Chen, J. (2011) The role of the human SWI5-MEI5 complex in homologous recombination repair. *Journal of Biological Chemistry*, **286**, 9888–9893.

108. Su, G.-C., Chung, C.-I., Liao, C.-Y., Lin, S.-W., Tsai, C.-T., Huang, T., Li, H.-W. and Chi, P. (2013) Enhancement of ADP release from the RAD51 presynaptic filament by the SWI5-SFR1 complex. *Nucleic Acids Research*, 10.1093/nar/gkt879.

109. Suwaki, N., Klare, K. and Tarsounas, M. (2011) RAD51 paralogs: roles in DNA damage signalling, recombinational repair and tumorigenesis. *Seminars in Cell and Developmental Biology*, **22**, 898–905.

110. Feng, Z., Scott, S.P., Bussen, W., Sharma, G.G., Guo, G., Pandita, T.K. and Powell, S.N. (2011) Rad52 inactivation is synthetically lethal with BRCA2 deficiency. *Proceedings of the National Academy of Sciences*, **108**, 686–691.

111. Mazin, A.V., Mazina, O.M., Bugreev, D.V. and Rossi, M.J. (2010) Rad54, the motor of homologous recombination. *DNA Repair*, **9**, 286–302.

112. Modesti, M., Budzowska, M., Baldeyron, C., Demmers, J.A.A., Ghirlando, R. and Kanaar, R. (2007) RAD51AP1 is a structure-specific DNA binding protein that stimulates joint molecule formation during RAD51-mediated homologous recombination. *Molecular Cell*, **28**, 468–481.

113. Wiese, C., Dray, E., Groesser, T., San Filippo, J., Shi, I., Collins, D.W., Tsai, M.-S., Williams, G.J., Rydberg, B., Sung, P., et al. (2007) Promotion of homologous recombination and genomic stability by RAD51AP1 via RAD51 recombinase enhancement. *Molecular Cell*, **28**, 482–490.

114. Richard, D.J., Bolderson, E., Cubeddu, L., Wadsworth, R.I.M., Savage, K., Sharma, G.G., Nicolette, M.L., Tsvetanov, S., McIlwraith, M.J., Pandita, R.K., et al. (2008) Single-stranded DNA-binding protein hSSB1 is critical for genomic stability. *Nature*, **453**, 677–681.

115. Li, Y., Bolderson, E., Kumar, R., Muniandy, P.A., Xue, Y., Richard, D.J., Seidman, M., Pandita, T.K., Khanna, K.K. and Wang, W. (2009) HSSB1 and hSSB2 form similar multiprotein complexes that participate in DNA damage response. *J. Biol. Chem.*, **284**, 23525–23531.

116. Moldovan, G.-L., Dejsuphong, D., Petalcorin, M.I.R., Hofmann, K., Takeda, S., Boulton, S.J. and D'Andrea, A.D. (2012) Inhibition of homologous recombination by the PCNA-interacting protein PARI. *Molecular Cell*, **45**, 75–86.

117. O'Connor, K.W., Dejsuphong, D., Park, E., Nicolae, C.M., Kimmelman, A.C., D'Andrea, A.D. and Moldovan, G.-L. (2013) PARI overexpression promotes genomic instability and pancreatic tumorigenesis. *Cancer Research*, **73**, 2529–2539.

118. Hu, Y., Raynard, S., Sehorn, M.G., Lu, X., Bussen, W., Zheng, L., Stark, J.M., Barnes, E.L., Chi, P., Janscak, P., et al. (2007) RECQL5/Recql5 helicase regulates homologous recombination and suppresses tumor formation via disruption of Rad51 presynaptic filaments. *Genes & Development*, **21**, 3073–3084.

119. Fugger, K., Mistrik, M., Danielsen, J.R., Dinant, C., Falck, J., Bartek, J., Lukas, J. and Mailand, N. (2009) Human Fbh1 helicase contributes to genome maintenance via pro- and anti-recombinase activities. *The Journal of Cell Biology*, **186**, 655–663.

120. Bacquin, A., Pouvelle, C., Siaud, N., Perderiset, M., Salomé-Desnoulez, S., Tellier-Lebegue, C., Lopez, B., Charbonnier, J.-B. and Kannouche, P.L. (2013) The helicase FBH1 is tightly regulated by PCNA via CRL4(Cdt2)-mediated proteolysis in human cells. *Nucleic Acids Research*, **41**, 6501–6513.

121. Barber, L.J., Youds, J.L., Ward, J.D., McIlwraith, M.J., O'Neil, N.J., Petalcorin, M.I.R., Martin, J.S., Collis, S.J., Cantor, S.B., Auclair, M., et al. (2008) RTEL1 maintains genomic stability by suppressing homologous recombination. *Cell*, **135**, 261–271.

122. Bachrati, C.Z., Borts, R.H. and Hickson, I.D. (2006) Mobile D-loops are a preferred substrate for the Bloom's syndrome helicase. *Nucleic Acids Research*, **34**, 2269–2279.

123. Chang, J.H., Kim, J.J., Choi, J.M., Lee, J.H. and Cho, Y. (2008) Crystal structure of the Mus81-Eme1 complex. *Genes & Development*, **22**, 1093–1106.

124. Ciccia, A., McDonald, N. and West, S.C. (2008) Structural and functional relationships of the XPF/MUS81 family of proteins. *Annu. Rev. Biochem.*, **77**, 259–287.

125. Xu, D., Guo, R., Sobeck, A., Bachrati, C.Z., Yang, J., Enomoto, T., Brown, G.W., Hoatlin, M.E., Hickson, I.D. and Wang, W. (2008) RMI, a new OB-fold complex essential for Bloom syndrome protein to maintain genome stability. *Genes & Development*, **22**, 2843–2855.

126. Rass, U., Compton, S.A., Matos, J., Singleton, M.R., Ip, S.C.Y., Blanco, M.G., Griffith, J.D. and West, S.C. (2010) Mechanism of Holliday junction resolution by the human GEN1 protein. *Genes & Development*, **24**, 1559–1569.

127. Fekairi, S., Scaglione, S., Chahwan, C., Taylor, E.R., Tissier, A., Coulon, S., Dong, M.-Q., Ruse, C., Yates, J.R., Russell, P., et al. (2009) Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases. *Cell*, **138**, 78–89.

128. Bussen, W., Raynard, S., Busygina, V., Singh, A.K. and Sung, P. (2007) Holliday junction processing activity of the BLM-Topo IIIalpha-BLAP75 complex. *J. Biol. Chem.*, **282**, 31484–31492.

129. Yang, J., Bachrati, C.Z., Ou, J., Hickson, I.D. and Brown, G.W. (2010) Human topoisomerase IIIalpha is a single-stranded DNA decatenase that is stimulated by BLM and RMI1. *Journal of Biological Chemistry*, **285**, 21426–21436.

130. Wyatt, H.D.M., Sarbajna, S., Matos, J. and West, S.C. (2013) Coordinated Actions of SLX1-SLX4 and MUS81-EME1 for Holliday Junction Resolution in Human Cells. *Molecular Cell*, 10.1016/j.molcel.2013.08.035.

131. Garner, E., Kim, Y., Lach, F.P., Kottemann, M.C. and Smogorzewska, A. (2013) Human GEN1 and the SLX4-Associated Nucleases MUS81 and SLX1 Are Essential for the Resolution of Replication-Induced Holliday Junctions. *Cell Rep*, 10.1016/j.celrep.2013.08.041.

132. Stark, J.M., Pierce, A.J., Oh, J., Pastink, A. and Jasin, M. (2004) Genetic steps of mammalian homologous repair with distinct mutagenic consequences. *Molecular and Cellular Biology*, **24**, 9305–9316.

133. Grimme, J.M., Honda, M., Wright, R., Okuno, Y., Rothenberg, E., Mazin, A.V., Ha, T. and Spies, M. (2010) Human Rad52 binds and wraps single-stranded DNA and mediates annealing via two hRad52-ssDNA complexes. *Nucleic Acids Research*, **38**, 2917–2930.

134. Al-Minawi, A.Z., Saleh-Gohari, N. and Helleday, T. (2008) The ERCC1/XPF endonuclease is required for efficient single-strand annealing and gene conversion in mammalian cells. *Nucleic Acids Research*, **36**, 1–9.

135. Göttlich, B., Reichenberger, S., Feldmann, E. and Pfeiffer, P. (1998) Rejoining of DNA double-strand breaks in vitro by single-strand annealing. *Eur. J. Biochem.*, **258**, 387–395.

136. Cortizas, E.M., Zahn, A., Hajjar, M.E., Patenaude, A.-M., Di Noia, J.M. and Verdun, R.E. (2013) Alternative End-Joining and Classical Nonhomologous End-Joining Pathways Repair Different Types of Double-Strand Breaks during Class-Switch Recombination. *J. Immunol.*, **191**, 5751–5763.

137. Boboila, C., Alt, F.W. and Schwer, B. (2012) Classical and alternative end-joining pathways for repair of lymphocyte-specific and general DNA double-strand breaks. *Adv. Immunol.*, **116**, 1–49.

138. Davis, A.J. and Chen, D.J. (2013) DNA double strand break repair via non-homologous end-joining. *Transl Cancer Res*, **2**, 130–143.

139. Lieber, M.R. (2010) The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.*, **79**, 181–211.

140. Mladenov, E. and Iliakis, G. (2011) Induction and repair of DNA double strand breaks: the increasing spectrum of non-homologous end joining pathways. *Mutat Res*, **711**, 61–72.

141. Walker, J.R., Corpina, R.A. and Goldberg, J. (2001) Structure of the Ku heterodimer bound to DNA and its implications for double-strand break repair. *Nature*, **412**, 607–614.

142. Uematsu, N., Weterings, E., Yano, K.-I., Morotomi-Yano, K., Jakob, B., Taucher-Scholz, G., Mari, P.-O., van Gent, D.C., Chen, B.P.C. and Chen, D.J. (2007) Autophosphorylation of DNA-PKCS regulates its dynamics at DNA double-strand breaks. *The Journal of Cell Biology*, **177**, 219–229.

143. Spagnolo, L., Rivera-Calzada, A., Pearl, L.H. and Llorca, O. (2006) Three-dimensional structure of the human DNA-PKcs/Ku70/Ku80 complex assembled on DNA and its implications for DNA DSB repair. *Molecular Cell*, **22**, 511–519.

144. Soubeyrand, S., Pope, L., De Chasseval, R., Gosselin, D., Dong, F., de Villartay, J.-P. and Haché, R.J.G. (2006) Artemis phosphorylated by DNA-dependent protein kinase associates preferentially with discrete regions of chromatin. *Journal of Molecular Biology*, **358**, 1200–1211.

145. Nick McElhinny, S.A., Snowden, C.M., McCarville, J. and Ramsden, D.A. (2000) Ku recruits the XRCC4-ligase IV complex to DNA ends. *Molecular and Cellular Biology*, **20**, 2996–3003.

146. Mari, P.-O., Florea, B.I., Persengiev, S.P., Verkaik, N.S., Brüggenwirth, H.T., Modesti, M., Giglia-Mari, G.,

Bezstarosti, K., Demmers, J.A.A., Luider, T.M., et al. (2006) Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 18597–18602.

147. Malivert, L., Ropars, V., Nunez, M., Drevet, P., Miron, S., Faure, G., Guerois, R., Mornon, J.-P., Revy, P., Charbonnier, J.-B., et al. (2010) Delineation of the Xrcc4-interacting region in the globular head domain of cernunnos/XLF. *Journal of Biological Chemistry*, **285**, 26475–26483.

148. Bernstein, N.K., Williams, R.S., Rakovszky, M.L., Cui, D., Green, R., Karimi-Busheri, F., Mani, R.S., Galicia, S., Koch, C.A., Cass, C.E., et al. (2005) The molecular architecture of the mammalian DNA repair enzyme, polynucleotide kinase. *Molecular Cell*, **17**, 657–670.

149. Ahel, I., Rass, U., El-Khamisy, S.F., Katyal, S., Clements, P.M., McKinnon, P.J., Caldecott, K.W. and West, S.C. (2006) The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. *Nature*, **443**, 713–716.

150. Roberts, S.A., Strande, N., Burkhalter, M.D., Strom, C., Havener, J.M., Hasty, P. and Ramsden, D.A. (2010) Ku is a 5'-dRP/AP lyase that excises nucleotide damage near broken ends. *Nature*, **464**, 1214–1217.

151. Li, S., Kanno, S.-I., Watanabe, R., Ogiwara, H., Kohno, T., Watanabe, G., Yasui, A. and Lieber, M.R. (2011) Polynucleotide kinase and aprataxin-like forkhead-associated protein (PALF) acts as both a single-stranded DNA endonuclease and a single-stranded DNA 3' exonuclease and can participate in DNA end joining in a biochemical system. *Journal of Biological Chemistry*, **286**, 36368–36377.

152. Ma, Y., Pannicke, U., Schwarz, K. and Lieber, M.R. (2002) Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell*, **108**, 781–794.

153. Povirk, L.F., Zhou, T., Zhou, R., Cowan, M.J. and Yannone, S.M. (2007) Processing of 3'-phosphoglycolate-terminated DNA double strand breaks by Artemis nuclease. *J. Biol. Chem.*, **282**, 3547–3558.

154. Perry, J.J.P., Yannone, S.M., Holden, L.G., Hitomi, C., Asaithamby, A., Han, S., Cooper, P.K., Chen, D.J. and Tainer, J.A. (2006) WRN exonuclease structure and molecular mechanism imply an editing role in DNA end processing. *Nature Structural & Molecular Biology*, **13**, 414–422.

155. Grundy, G.J., Rulten, S.L., Zeng, Z., Arribas-Bosacoma, R., Iles, N., Manley, K., Oliver, A. and Caldecott, K.W. (2013) APLF promotes the assembly and activity of non-homologous end joining protein complexes. *EMBO J*, **32**, 112–125.

156. Ma, Y., Lu, H., Tippin, B., Goodman, M.F., Shimazaki, N., Koiwai, O., Hsieh, C.-L., Schwarz, K. and Lieber, M.R. (2004) A biochemically defined system for mammalian nonhomologous DNA end joining. *Molecular Cell*, **16**, 701–713.

157. Dueva, R. and Iliakis, G. (2013) Alternative pathways of non-homologous end joining (NHEJ) in genomic instability and cancer. *Transl Cancer Res*, **2**, 163–177.

158. Truong, L.N., Li, Y., Shi, L.Z., Hwang, P.Y.-H., He, J., Wang, H., Razavian, N., Berns, M.W. and Wu, X. (2013) Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proceedings of the National Academy of Sciences*, **110**, 7720–7725.

159. Mansour, W.Y., Rhein, T. and Dahm-Daphi, J. (2010) The alternative end-joining pathway for repair of DNA double-strand breaks requires PARP1 but is not dependent upon microhomologies. *Nucleic Acids Research*, **38**, 6065–6077.

160. Wang, M., Wu, W., Wu, W., Rosidi, B., Zhang, L., Wang, H. and Iliakis, G. (2006) PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways. *Nucleic Acids Research*, **34**, 6170–6182.

161. Cheng, Q., Barboule, N., Frit, P., Gomez, D., Bombarde, O., Couderc, B., Ren, G.S., Salles, B. and Calsou, P. (2011) Ku counteracts mobilization of PARP1 and MRN in chromatin damaged with DNA double-strand breaks. *Nucleic Acids Research*, **39**, 9605–9619.

162. Zhong, Q., Chen, C.-F., Chen, P.-L. and Lee, W.-H. (2002) BRCA1 facilitates microhomology-mediated end

joining of DNA double strand breaks. *J. Biol. Chem.*, **277**, 28641–28647.

163. Lee-Theilen, M., Matthews, A.J., Kelly, D., Zheng, S. and Chaudhuri, J. (2011) CtIP promotes microhomology-mediated alternative end joining during class-switch recombination. *Nature Structural & Molecular Biology*, **18**, 75–79.

164. Zhang, Y. and Jasin, M. (2011) An essential role for CtIP in chromosomal translocation formation through an alternative end-joining pathway. *Nature Structural & Molecular Biology*, **18**, 80–84.

165. Xie, A., Kwok, A. and Scully, R. (2009) Role of mammalian Mre11 in classical and alternative nonhomologous end joining. *Nature Structural & Molecular Biology*, **16**, 814–818.

166. Rass, E., Grabarz, A., Plo, I., Gautier, J., Bertrand, P. and Lopez, B.S. (2009) Role of Mre11 in chromosomal nonhomologous end joining in mammalian cells. *Nature Structural & Molecular Biology*, **16**, 819–824.

167. Bothmer, A., Robbiani, D.F., Feldhahn, N., Gazumyan, A., Nussenzweig, A. and Nussenzweig, M.C. (2010) 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. *Journal of Experimental Medicine*, **207**, 855–865.

168. Bothmer, A., Robbiani, D.F., Di Virgilio, M., Bunting, S.F., Klein, I.A., Feldhahn, N., Barlow, J., Chen, H.-T., Bosque, D., Callen, E., et al. (2011) Regulation of DNA end joining, resection, and immunoglobulin class switch recombination by 53BP1. *Molecular Cell*, **42**, 319–329.

169. Wang, H., Rosidi, B., Perrault, R., Wang, M., Zhang, L., Windhofer, F. and Iliakis, G. (2005) DNA ligase III as a candidate component of backup pathways of nonhomologous end joining. *Cancer Research*, **65**, 4020–4030.

170. Liang, L., Deng, L., Nguyen, S.C., Zhao, X., Maulion, C.D., Shao, C. and Tischfield, J.A. (2008) Human DNA ligases I and III, but not ligase IV, are required for microhomology-mediated end joining of DNA double-strand breaks. *Nucleic Acids Research*, **36**, 3297–3310.

171. Della-Maria, J., Zhou, Y., Tsai, M.-S., Kuhnlein, J., Carney, J.P., Paull, T.T. and Tomkinson, A.E. (2011) Human Mre11/human Rad50/Nbs1 and DNA ligase IIIalpha/XRCC1 protein complexes act together in an alternative nonhomologous end joining pathway. *Journal of Biological Chemistry*, **286**, 33845–33853.

172. Paul, K., Wang, M., Mladenov, E., Bencsik-Theilen, A., Bednar, T., Wu, W., Arakawa, H. and Iliakis, G. (2013) DNA ligases I and III cooperate in alternative non-homologous end-joining in vertebrates. *PLoS ONE*, **8**, e59505.

173. Zeman, M.K. and Cimprich, K.A. (2014) Causes and consequences of replication stress. *Nat. Cell Biol.*, **16**, 2–9.

174. Branzei, D. and Foiani, M. (2010) Maintaining genome stability at the replication fork. *Nature Publishing Group*, **11**, 208–219.

175. Hanada, K., Budzowska, M., Davies, S.L., van Drunen, E., Onizawa, H., Beverloo, H.B., Maas, A., Essers, J., Hickson, I.D. and Kanaar, R. (2007) The structure-specific endonuclease Mus81 contributes to replication restart by generating double-strand DNA breaks. *Nature Structural & Molecular Biology*, **14**, 1096–1104.

176. Chen, S. and Parmigiani, G. (2007) Meta-analysis of BRCA1 and BRCA2 penetrance. *J. Clin. Oncol.*, **25**, 1329–1333.

177. Howlader, N., Noone, A.M., Krapcho, M., Garshell, J., Neyman, N., Altekruse, S.F., Kosary, C.L., Yu, M., Ruhl, J., Tatalovich, Z., et al. SEER Cancer Statistics Review, 1975-2010. *seer.cancer.gov*.

178. Cerbinskaite, A., Mukhopadhyay, A., Plummer, E.R., Curtin, N.J. and Edmondson, R.J. (2012) Defective homologous recombination in human cancers. *Cancer Treatment Reviews*, **38**, 89–100.

179. Khanna, K.K. and Chenevix-Trench, G. (2004) ATM and genome maintenance: defining its role in breast cancer susceptibility. *J Mammary Gland Biol Neoplasia*, **9**, 247–262.

180. Swift, M., Reitnauer, P.J., Morrell, D. and Chase, C.L. (1987) Breast and other cancers in families with ataxia-telangiectasia. *N. Engl. J. Med.*, **316**, 1289–1294.

181. Jones, S., Zhang, X., Parsons, D.W., Lin, J.C.-H., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama,

H., Jimeno, A., et al. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, **321**, 1801–1806.

182. Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

183. Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 10.1038/nature10983.

184. Shah, S.P., Roth, A., Goya, R., Oloumi, A., Ha, G., Zhao, Y., Turashvili, G., Ding, J., Tse, K., Haffari, G., et al. (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**, 395–399.

185. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature*, **486**, 400–404.

186. Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L., et al. (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, **486**, 405–409.

187. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M.R. (2004) A census of human cancer genes. *Nat Rev Cancer*, **4**, 177–183.

188. Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell*, **144**, 27–40.

189. Yang, L., Luquette, L.J., Gehlenborg, N., Xi, R., Haseley, P.S., Hsieh, C.-H., Zhang, C., Ren, X., Protopopov, A., Chin, L., et al. (2013) Diverse mechanisms of somatic structural variations in human cancer genomes. *Cell*, **153**, 919–929.

190. Zhang, C.-Z., Leibowitz, M.L. and Pellman, D. (2013) Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. *Genes & Development*, **27**, 2513–2530.

191. Iwakawa, R., Takenaka, M., Kohno, T., Shimada, Y., Totoki, Y., Shibata, T., Tsuta, K., Nishikawa, R., Noguchi, M., Sato-Otsubo, A., et al. (2013) Genome-wide identification of genes with amplification and/or fusion in small cell lung cancer. *Genes Chromosomes Cancer*, **52**, 802–816.

192. Malhotra, A., Lindberg, M., Faust, G.G., Leibowitz, M.L., Clark, R.A., Layer, R.M., Quinlan, A.R. and Hall, I.M. (2013) Breakpoint profiling of 64 cancer genomes reveals numerous complex rearrangements spawned by homology-independent mechanisms. *Genome Research*, **23**, 762–776.

193. Rausch, T., Jones, D.T.W., Zapatka, M., Stütz, A.M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P.A., et al. (2012) Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. *Cell*, **148**, 59–71.

194. Kloosterman, W.P., Guryev, V., van Roosmalen, M., Duran, K.J., de Bruijn, E., Bakker, S.C.M., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M., et al. (2011) Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Human Molecular Genetics*, **20**, 1916–1924.

195. Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415–421.

196. Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N. and Sander, C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, **45**, 1127–1133.

197. Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*,

149, 979–993.

198. Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al. (2012) The life history of 21 breast cancers. *Cell*, **149**, 994–1007.

199. Koito, A. and Ikeda, T. (2013) Intrinsic immunity against retrotransposons by APOBEC cytidine deaminases. *Front Microbiol*, **4**, 28.

200. Burns, M.B., Lackey, L., Carpenter, M.A., Rathore, A., Land, A.M., Leonard, B., Refsland, E.W., Kotandeniya, D., Tretyakova, N., Nikas, J.B., et al. (2013) APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature*, **494**, 366–370.

201. Al-Ejeh, F., Kumar, R., Wiegmans, A., Lakhani, S.R., Brown, M.P. and Khanna, K.K. (2010) Harnessing the complexity of DNA-damage response pathways to improve cancer treatment outcomes. *Oncogene*, **29**, 6085–6098.

202. Evers, B., Helleday, T. and Jonkers, J. (2010) Targeting homologous recombination repair defects in cancer. *Trends Pharmacol. Sci.*, **31**, 372–380.

203. Bolderson, E., Richard, D.J., Zhou, B.-B.S. and Khanna, K.K. (2009) Recent advances in cancer therapy targeting proteins involved in DNA double-strand break repair. *Clin. Cancer Res.*, **15**, 6314–6320.

204. Cheung-Ong, K., Giaever, G. and Nislow, C. (2013) DNA-damaging agents in cancer chemotherapy: serendipity and chemical biology. *Chem. Biol.*, **20**, 648–659.

205. Boone, C., Bussey, H. and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, **8**, 437–449.

206. Nijman, S.M.B. (2011) Synthetic lethality: general principles, utility and detection using genetic screens in human cells. *FEBS Letters*, **585**, 1–6.

207. O'Neil, N.J., van Pel, D.M. and Hieter, P. (2013) Synthetic lethality and cancer: cohesin and PARP at the replication fork. *Trends Genet.*, **29**, 290–297.

208. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

209. Sandmann, T. and Boutros, M. (2012) Screens, maps & networks: from genome sequences to personalized medicine. *Curr. Opin. Genet. Dev.*, **22**, 36–44.

210. Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., et al. (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**, 917–921.

211. Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J. and Helleday, T. (2005) Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, **434**, 913–917.

212. Fong, P.C., Boss, D.S., Yap, T.A., Tutt, A., Wu, P., Mergui-Roelvink, M., Mortimer, P., Swaisland, H., Lau, A., O'Connor, M.J., et al. (2009) Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *N. Engl. J. Med.*, **361**, 123–134.

213. Tsutsumi, M., Masutani, M., Nozaki, T., Kusuoka, O., Tsujiuchi, T., Nakagama, H., Suzuki, H., Konishi, Y. and Sugimura, T. (2001) Increased susceptibility of poly(ADP-ribose) polymerase-1 knockout mice to nitrosamine carcinogenicity. *Carcinogenesis*, **22**, 1–3.

214. Shall, S. and de Murcia, G. (2000) Poly(ADP-ribose) polymerase-1: what have we learned from the deficient mouse model? *Mutat Res*, **460**, 1–15.

215. Gottipati, P., Vischioni, B., Schultz, N., Solomons, J., Bryant, H.E., Djureinovic, T., Issaeva, N., Sleeth, K., Sharma, R.A. and Helleday, T. (2010) Poly(ADP-ribose) polymerase is hyperactivated in homologous recombination-defective cells. *Cancer Research*, **70**, 5389–5398.

216. Patel, A.G., Sarkaria, J.N. and Kaufmann, S.H. (2011) Nonhomologous end joining drives poly(ADP-ribose)

polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells. *Proceedings of the National Academy of Sciences*, **108**, 3406–3411.

217. Helleday, T. (2011) The underlying mechanism for the PARP and BRCA synthetic lethality: Clearing up the misunderstandings. *Molecular Oncology*, **5**, 387–393.

218. Yang, Y.-G., Cortes, U., Patnaik, S., Jasin, M. and Wang, Z.-Q. (2004) Ablation of PARP-1 does not interfere with the repair of DNA double-strand breaks, but compromises the reactivation of stalled replication forks. *Oncogene*, **23**, 3872–3882.

219. Ying, S., Hamdy, F.C. and Helleday, T. (2012) Mre11-dependent degradation of stalled DNA replication forks is prevented by BRCA2 and PARP1. *Cancer Research*, **72**, 2814–2821.

220. Bryant, H.E., Petermann, E., Schultz, N., Jemth, A.-S., Loseva, O., Issaeva, N., Johansson, F., Fernandez, S., McGlynn, P. and Helleday, T. (2009) PARP is activated at stalled forks to mediate Mre11-dependent replication restart and recombination. *EMBO J*, **28**, 2601–2615.

221. Ström, C. and Helleday, T. (2012) Strategies for the Use of Poly(adenosine diphosphate ribose) Polymerase (PARP) Inhibitors in Cancer Therapy. *Biomolecules*, **2**, 635–649.

222. Curtin, N.J. (2005) PARP inhibitors for cancer therapy. *Expert Rev Mol Med*, **7**, 1–20.

223. Murai, J., Huang, S.-Y.N., Das, B.B., Renaud, A., Zhang, Y., Doroshow, J.H., Ji, J., Takeda, S. and Pommier, Y. (2012) Trapping of PARP1 and PARP2 by Clinical PARP Inhibitors. *Cancer Research*, **72**, 5588–5599.

224. Bartkova, J., Rezaei, N., Liontos, M., Karakaidos, P., Kletsas, D., Issaeva, N., Vassiliou, L.-V.F., Kolettas, E., Niforou, K., Zoumpourlis, V.C., et al. (2006) Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature*, **444**, 633–637.

225. de Murcia, J.M., Niedergang, C., Trucco, C., Ricoul, M., Dutrillaux, B., Mark, M., Oliver, F.J., Masson, M., Dierich, A., LeMeur, M., et al. (1997) Requirement of poly(ADP-ribose) polymerase in recovery from DNA damage in mice and in cells. *Proc. Natl. Acad. Sci. U.S.A.*, **94**, 7303–7307.

226. Wang, Z.Q., Stingl, L., Morrison, C., Jantsch, M., Los, M., Schulze-Osthoff, K. and Wagner, E.F. (1997) PARP is important for genomic stability but dispensable in apoptosis. *Genes & Development*, **11**, 2347–2358.

227. McCabe, N., Turner, N.C., Lord, C.J., Kluzek, K., Bialkowska, A., Swift, S., Giavara, S., O'Connor, M.J., Tutt, A.N., Zdzienicka, M.Z., et al. (2006) Deficiency in the repair of DNA damage by homologous recombination and sensitivity to poly(ADP-ribose) polymerase inhibition. *Cancer Research*, **66**, 8109–8115.

228. Drost, R., Bouwman, P., Rottenberg, S., Boon, U., Schut, E., Klarenbeek, S., Klijn, C., van der Heijden, I., van der Gulden, H., Wientjens, E., et al. (2011) BRCA1 RING function is essential for tumor suppression but dispensable for therapy resistance. *Cancer Cell*, **20**, 797–809.

229. Drost, R. and Jonkers, J. (2013) Opportunities and hurdles in the treatment of BRCA1-related breast cancer. *Oncogene*, 10.1038/onc.2013.329.

230. Swisher, E.M., Sakai, W., Karlan, B.Y., Wurz, K., Urban, N. and Taniguchi, T. (2008) Secondary BRCA1 mutations in BRCA1-mutated ovarian carcinomas with platinum resistance. *Cancer Research*, **68**, 2581–2586.

231. Norquist, B., Wurz, K.A., Pennil, C.C., Garcia, R., Gross, J., Sakai, W., Karlan, B.Y., Taniguchi, T. and Swisher, E.M. (2011) Secondary somatic mutations restoring BRCA1/2 predict chemotherapy resistance in hereditary ovarian carcinomas. *J. Clin. Oncol.*, **29**, 3008–3015.

232. Bouwman, P. and Jonkers, J. (2012) The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. *Nat Rev Cancer*, **12**, 587–598.

233. Edwards, S.L., Brough, R., Lord, C.J., Natrajan, R., Vatcheva, R., Levine, D.A., Boyd, J., Reis-Filho, J.S. and Ashworth, A. (2008) Resistance to therapy caused by intragenic deletion in BRCA2. *Nature*, **451**, 1111–1115.

234. Sakai, W., Swisher, E.M., Karlan, B.Y., Agarwal, M.K., Higgins, J., Friedman, C., Villegas, E., Jacquemont, C., Farrugia, D.J., Couch, F.J., et al. (2008) Secondary mutations as a mechanism of cisplatin resistance in BRCA2-

mutated cancers. *Nature*, **451**, 1116–1120.

235. Jaspers, J.E., Kersbergen, A., Boon, U., Sol, W., van Deemter, L., Zander, S.A., Drost, R., Wientjens, E., Ji, J., Aly, A., et al. (2013) Loss of 53BP1 causes PARP inhibitor resistance in Brca1-mutated mouse mammary tumors. *Cancer Discov*, **3**, 68–81.

236. Johnson, N., Johnson, S.F., Yao, W., Li, Y.-C., Choi, Y.-E., Bernhardy, A.J., Wang, Y., Capelletti, M., Sarosiek, K.A., Moreau, L.A., et al. (2013) Stabilization of mutant BRCA1 protein confers PARP inhibitor and platinum resistance. *Proceedings of the National Academy of Sciences*, **110**, 17041–17046.

237. Pace, P., Mosedale, G., Hodskinson, M.R., Rosado, I.V., Sivasubramaniam, M. and Patel, K.J. (2010) Ku70 corrupts DNA repair in the absence of the Fanconi anemia pathway. *Science*, **329**, 219–223.

238. Adamo, A., Collis, S.J., Adelman, C.A., Silva, N., Horejsí, Z., Ward, J.D., Martinez-Perez, E., Boulton, S.J. and La Volpe, A. (2010) Preventing nonhomologous end joining suppresses DNA repair defects of Fanconi anemia. *Molecular Cell*, **39**, 25–35.

239. Barabási, A.-L., Gulbahce, N. and Loscalzo, J. (2011) Network medicine: a network-based approach to human disease. *Nat Rev Genet*, **12**, 56–68.

240. Michor, F., Hughes, T.P., Iwasa, Y., Branford, S., Shah, N.P., Sawyers, C.L. and Nowak, M.A. (2005) Dynamics of chronic myeloid leukaemia. *Nature*, **435**, 1267–1270.

241. Srihari, S. and Ragan, M.A. (2013) Systematic tracking of dysregulated modules identifies novel genes in cancer. *Bioinformatics*, **29**, 1553–1561.

242. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, **106**, 9362–9367.

243. Li,F.P., Fraumeni,J.F., Mulvihill,J.J., Blattner,W.A., Dreyfus,M.G., Tucker,M.A. and Miller,R.W. (1988) A cancer family syndrome in twenty-four kindreds. *Cancer Research*, **48**, 5358–5362.

244. Masciari,S., Dillon,D.A., Rath,M., Robson,M., Weitzel,J.N., Balmana,J., Gruber,S.B., Ford,J.M., Euhus,D., Lebensohn,A., et al. (2012) Breast cancer phenotype in women with TP53 germline mutations: a Li-Fraumeni syndrome consortium effort. *Breast Cancer Res Treat*, **133**, 1125–1130.

245. Lee,J., Kim,D.-H., Lee,S., Yang,Q.-H., Lee,D.K., Lee,S.-K., Roeder,R.G. and Lee,J.W. (2009) A tumor suppressive coactivator complex of p53 containing ASC-2 and histone H3-lysine-4 methyltransferase MLL3 or its paralogue MLL4. *Proceedings of the National Academy of Sciences*, **106**, 8513–8518.

246. Bennetzen,M.V., Larsen,D.H., Dinant,C., Watanabe,S., Bartek,J., Lukas,J. and Andersen,J.S. (2013) Acetylation dynamics of human nuclear proteins during the ionizing radiation-induced DNA damage response. *Cell Cycle*, **12**, 1688–1695.

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

248. Liaw,D., Marsh,D.J., Li,J., Dahia,P.L., Wang,S.I., Zheng,Z., Bose,S., Call,K.M., Tsou,H.C., Peacocke,M., et al. (1997) Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat. Genet.*, **16**, 64–67.

249. Tavtigian,S.V., Oefner,P.J., Babikyan,D., Hartmann,A., Healey,S., Le Calvez-Kelm,F., Lesueur,F., Byrnes,G.B., Chuang,S.-C., Forey,N., et al. (2009) Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am. J. Hum. Genet.*, **85**, 427–446.

250. Vo,Q.N., Kim,W.-J., Cvitanovic,L., Boudreau,D.A., Ginzinger,D.G. and Brown,K.D. (2004) The ATM gene is a target for epigenetic silencing in locally advanced breast cancer. *Oncogene*, **23**, 9432–9437.

251. Brennan,K., García-Closas,M., Orr,N., Fletcher,O., Jones,M., Ashworth,A., Swerdlow,A., Thorne,H., kConFab Investigators, Riboli,E., et al. (2012) Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. *Cancer Research*, **72**, 2304–2313.

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

253. Esteller,M., Silva,J.M., Dominguez,G., Bonilla,F., Matias-Guiu,X., Lerma,E., Bussaglia,E., Prat,J., Harkes,I.C., Repasky,E.A., et al. (2000) Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *J. Natl. Cancer Inst.*, **92**, 564–569.

254. Puc,J., Keniry,M., Li,H.S., Pandita,T.K., Choudhury,A.D., Memeo,L., Mansukhani,M., Murty,V.V.V.S., Gaciong,Z., Meek,S.E.M., et al. (2005) Lack of PTEN sequesters CHK1 and initiates genetic instability. *Cancer Cell*, **7**, 193–204.

255. Marees,T., Moll,A.C., Imhof,S.M., de Boer,M.R., Ringens,P.J. and van Leeuwen,F.E. (2008) Risk of second malignancies in survivors of retinoblastoma: more than 40 years of follow-up. *J. Natl. Cancer Inst.*, **100**, 1771–1779.

256. Seal,S., Thompson,D., Renwick,A., Elliott,A., Kelly,P., Barfoot,R., Chagtai,T., Jayatilake,H., Ahmed,M., Spanova,K., et al. (2006) Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.*, **38**, 1239–1241.

257. 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. (2002) Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.*, **31**, 55–59.

258. Wiesner,T., Fried,I., Ulz,P., Stacher,E., Popper,H., Murali,R., Kutzner,H., Lax,S., Smolle-Jüttner,F., Geigl,J.B., et al. (2012) Toward an improved definition of the tumor spectrum associated with BAP1 germline mutations. *J. Clin. Oncol.*, **30**, e337–40.

259. Jirawatnotai,S., Hu,Y., Michowski,W., Elias,J.E., Becks,L., Bienvenu,F., Zagozdzon,A., Goswami,T., Wang,Y.E., Clark,A.B., et al. (2011) A function for cyclin D1 in DNA repair uncovered by protein interactome analyses in human cancers. *Nature*, **474**, 230–234.

260. French,J.D., Ghoussaini,M., Edwards,S.L., Meyer,K.B., Michailidou,K., Ahmed,S., Khan,S., Maranian,M.J., O'Reilly,M., Hillman,K.M., et al. (2013) Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.*, **92**, 489–503.

261. Rahman,N., Seal,S., Thompson,D., Kelly,P., Renwick,A., Elliott,A., Reid,S., Spanova,K., Barfoot,R., Chagtai,T., et al. (2007) PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.*, **39**, 165–167.

Chapter Three: Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer

Compared to other DNA repair pathways, HR is the most complicated one, which is also the most closely related to breast cancer susceptibility and therapy response. Understanding the cellular status of HR repair is thus of key importance for breast cancer research and treatment. However, HR deficiency can be caused by various mechanisms, including loss-of-function mutations in key genes, inappropriate post-translational modifications and defects in chromatin remodelling, which make it difficult to directly detect HR status by interrogating one gene or one mechanism.

In this chapter, I calculated an expression-based HR score to quantify HR pathway dysregulation in individual sporadic breast tumours. The results based on this score show that the degree of HR dysregulation varies from sample to sample, and samples with highly dysregulated HR are very likely to be HR deficient. More importantly, I uncovered a novel association between HR dysregulation and chromosomal instability (CIN), indicating that compromised HR activity might be an important cause of the CIN observed in sporadic breast cancer. This result helps pinpoint the molecular basis of CIN in sporadic cancers, which remains poorly understood so far, and has important implications for understanding CIN-related tumour evolution and drug resistance.

The Pathifier method used to calculate the HR score is available as an R package (Drier et al., 2013). The working principle of this method has been briefly described in Chapter One (Section 1.4.4 Page 19) and also in the publication below. To be more specific, Pathifier evaluates one pathway at a time, calculating for each sample a pathway deregulation score (PDS) that quantifies the extent to which the pathway is dysregulated in a particular tumour sample. Pathifier requires as input the expression levels of the component genes of a given pathway in both tumour and normal samples, with the latter serving as a benchmark.

The calculation of a PDS is a four-step process. First, the absolute expression values are normalised to account for differences in variation between genes. Next is a principal component analysis (PCA) to reduce dimensionality of the input data and counteract the effects caused by noisy or highly correlated genes. In the following step, Pathifier adopts the Hastie and Stuetzle's algorithm (Hastie and Stuetzle, 1989) to construct a principal curve that best describes the variability of the entire sample set. Lastly each sample, including normal and tumour, is projected onto the principal curve;

the PDS of a tumour sample is defined as its distance along the curve from the centroid of the projections of the normal samples. More technical details concerning these steps can be found in the publications by the Pathifier authors (Drier et al., 2013; Livshits et al., 2015).

Although Pathifier is emerging as a promising method for personalised pathway analysis, it harbours a limitation that, while each sample can be assigned a PDS for a particular pathway, the calculation of the PDS requires the entire sample set to build the principal curve, as described above. Therefore, this method cannot be used in studies where there are only a few samples. Besides, the PDSs of samples from different cohorts are not directly comparable as each score is calculated in the context of a particular sample set (i.e., the PDSs are relative values rather than absolute values). However, these limitations do not affect the results presented in this chapter as they were obtained through the independent analyses of four different datasets.

These four datasets are from two major genomic studies in breast cancer in recent years, performed by the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) (Curtis et al., 2012) and The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Network, 2012), respectively. Descriptions about these datasets, including the number of samples in each breast cancer subtype, and how the respective gene expression data and DNA copy-number data had been generated and pre-processed, are provided in the Materials and Methods section of the following publication. Further details about these datasets can be found in the original METABRIC and TCGA publications (Cancer Genome Atlas Network, 2012; Curtis et al., 2012).

Results presented as a publication

The results were published in the peer-reviewed journal *Molecular Oncology* in September 2015. The supplementary figures of this publication are presented in Appendix 2, and the supplementary tables (which are all Excel files) are deposited at UQ eSpace.

References

Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. doi:10.1038/nature11412

Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al., The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346-352. doi:10.1038/nature10983

Drier, Y., Sheffer, M., Domany, E., 2013. Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences of the United States of America 110, 6388–6393. doi:10.1073/pnas.1219651110

Hastie, T., Stuetzle, W., 1989. Principal curves. Journal of the American Statistical Association 84, 502–516.

Livshits, A., Git, A., Fuks, G., Caldas, C., Domany, E., 2015. Pathway-based personalized analysis of breast cancer expression data. Molecular Oncology 9, 1471-1483. doi:10.1016/j.molonc.2015.04.006

Personalised Pathway Analysis Reveals Association between DNA Repair Pathway Dysregulation and Chromosomal Instability in Sporadic Breast Cancer

Chao Liu[a], Sriganesh Srihari[a], Samir Lal[b], Benoît Gautier[c], Peter T. Simpson[bd], Kum Kum Khanna[e], Mark A. Ragan[a*] and Kim-Anh Lê Cao[c*]

[a]Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4067, Australia
[b]The University of Queensland, UQ Centre for Clinical Research, Herston, QLD 4029, Australia
[c]University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD 4102, Australia
[d]School of Medicine, The University of Queensland, Herston, QLD 4006, Australia
[e]QIMR-Berghofer Medical Research Institute, Herston, Brisbane, QLD 4029, Australia

[*] Joint corresponding authors:
Dr. Kim-Anh Lê Cao, University of Queensland Diamantina Institute, Translational Research Institute, Level 5, 37 Kent Street, Woolloongabba, QLD 4102, Australia. Tel: +61 7 3443 7069; Fax: +61 7 3443 6966; Email: k.lecao@uq.edu.au

Prof Mark A. Ragan, Institute for Molecular Bioscience, The University of Queensland, Level 6, 306 Carmody Rd, St. Lucia, QLD 4067, Australia. Tel.: +61 7 3346 2616; Fax: +61 7 3346 2101; E-mail: m.ragan@uq.edu.au

Abstract

The Homologous Recombination (HR) pathway is crucial for the repair of DNA double-strand breaks (DSBs) generated during DNA replication. Defects in HR repair have been linked to the initiation and development of a wide variety of human malignancies, and exploited in chemical, radiological and targeted therapies. In this study, we performed a personalised pathway analysis independently for four large sporadic breast cancer cohorts to investigate the status of HR pathway dysregulation in individual sporadic breast tumours, its association with HR repair deficiency and its impact on tumour characteristics. Specifically, we first manually curated a list of HR genes according to our recent review on this pathway (Liu et al., 2014), and then applied a personalised pathway analysis method named Pathifier (Drier et al., 2013) on the expression levels of the curated genes to obtain an *HR* score quantifying HR pathway dysregulation in individual tumours. Based on the score, we observed a great diversity in HR dysregulation between and within gene expression-based breast cancer subtypes, and by using two published HR-defect signatures, we found HR pathway dysregulation reflects HR repair deficiency. Furthermore, we identified a novel association between HR pathway dysregulation and chromosomal instability (CIN) in sporadic breast cancer. Although CIN has long been considered as a hallmark of most solid tumours, with recent extensive studies highlighting its importance in tumour evolution and drug resistance, the molecular basis of CIN in sporadic cancers remains poorly understood. Our results imply that HR pathway dysregulation might contribute to CIN in sporadic breast cancer.

Keywords: *DNA repair*; *homologous recombination*; *breast cancer*; *chromosomal instability*; *pathway analysis*

Introduction

Chromosomal instability (CIN), defined as an increased rate of gain or loss of whole chromosomes or large chromosomal fragments, is a hallmark of most solid tumours. CIN is the primary form of genomic instability that is thought to be the major cause of genetic heterogeneity in cancer (Burrell et al., 2013b), and is thus strongly implicated in tumour evolution. CIN also has important clinical implications, as it has been linked to poor prognosis e.g. by conferring intrinsic multidrug resistance (Lee et al., 2011). The molecular basis of CIN in hereditary cancer is relatively clear, and has been attributed to mutations in DNA repair genes (Negrini et al., 2010); however, the underlying mechanisms of CIN in various sporadic cancers remain poorly understood. Carter and colleagues developed a gene expression-based CIN signature, termed CIN25, based on 25 genes that are most overexpressed in tumours with CIN (Carter et al., 2006). A considerable number of genes involved in *replication* and *cell cycle* contribute to this signature, suggesting an important link between these cellular processes and CIN. This was further corroborated by Negrini et al. (2010), who proposed a *replication stress* model to explain CIN in sporadic tumours; this model was recently validated in colorectal cancer (Burrell et al., 2013a).

Highly proliferative cancer cells undergo considerable replication stress that results in the stalling of replication forks. These stalled forks are usually stabilised and restarted after the source of stress is removed via a complex replication stress response pathway (Zeman and Cimprich, 2014). Lack of stabilisation and/or the prolonged persistence of a stalled fork can generate DNA double-strand breaks (DSBs), which are subsequently repaired by DSB repair machinery to restart the forks. However, in the absence of such a DSB repair machinery the DSBs will develop into chromosomal breaks, resulting in CIN. *Homologous recombination* (HR) is a crucial pathway responsible for repairing DSBs during replication. Using homologous sister chromatid as templates, HR presents a high-fidelity repair mechanism that is crucial for error-free DNA replication.

The core components of HR are fairly well established for their specific roles i.e. monitoring, signalling and repairing of DSBs (Liu et al., 2014), and HR defects can be detected by investigating the loss-of-function mutations in these genes. However, the dysfunction of HR can also be caused by numerous other mechanisms. For example, changes or defects in chromatin remodelling (Price and D'Andrea, 2013; van Attikum and Gasser, 2009), microRNAs (Chowdhury et al., 2013; d'Adda di Fagagna, 2014; Sharma and Misteli, 2013), post-translational modifications such as ubiquitination and sumoylation (Bekker-Jensen and Mailand, 2011; Dou et al., 2011; Ulrich, 2012), and inappropriate expression of certain genes that are not directly involved in HR (Y. Peng et al., 2015; Watkins et al., 2015) can considerably affect HR components, thereby causing aberrant HR function. As a consequence, single-gene approaches or approaches focussing

on one mechanism yield only an incomplete picture of abnormal HR in a given tumour. On the other hand, HR-deficient cells may compensate for the defect in a given HR gene by altering the expression level of other HR genes (Pitroda et al., 2014). The most notable example is the overexpression of DNA repair protein RAD51 homolog 1 (*RAD51*), which is observed when breast cancer susceptibility gene 1 (*BRCA1*) (Martin et al., 2007), breast cancer susceptibility gene 2 (*BRCA2*) (Brown and Holt, 2009) or other key HR genes (Takata et al., 2001) are defective. It is therefore of interest to determine a measure of HR *pathway* dysregulation, aggregating the expression of all HR genes, which may reflect HR repair deficiency in tumours regardless of the mechanism that has led to the deficiency.

The vast majority of breast tumours are sporadic, accounting for 90%-95% of all diagnosed breast cancer cases (Davis, 2011) and are characterised by their great heterogeneity in biological property and patient outcome. To dissect this heterogeneity, estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) have been used as standardised diagnostic markers in clinical practice to guide the choice of treatment. Gene expression profiling has defined five intrinsic subtypes (also known as PAM50 subtypes) with clinical relevance: Luminal A, Luminal B, Basal-like, HER2 and Normal-like (Hu et al., 2006; Parker et al., 2009; Perou et al., 2000; Sørlie et al., 2001). More-recent genomic studies, notably from the Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), have uncovered substantial heterogeneities within these receptor- or gene expression-based subtypes, resulting in the definition of up to ten subtypes (Ciriello et al., 2013; Curtis et al., 2012; Koboldt et al., 2012; Lehmann et al., 2011; Yanagawa et al., 2012). However, it is likely that heterogeneity exists even within these newly established subtypes. In the coming age of personalised medicine, each tumour may be analysed individually.

Pathway analysis has become the first choice to gain functional insights from expression data, beyond the detection of differential genes. Numerous pathway analysis tools have been developed, however, most of them are designed for providing pathway dysregulation information at population level instead of tumour level. Among the recently proposed methods for personalised pathways analysis (Ahn et al., 2014; Drier et al., 2013; Vaske et al., 2010; Wang et al., 2015a; 2015b), Pathifier (Drier et al., 2013) has proven to be particularly robust. It has been successfully applied to provide a pathway-based classification of breast cancer (Livshits et al., 2015), and when combined with Cox regression and L1 penalised estimation, has achieved better prognosis prediction compared with gene-based models (Huang et al., 2014).

In this study, we sought to perform a personalised pathway analysis to obtain a comprehensive understanding of the status of HR pathway dysregulation in individual sporadic breast tumours, its association with HR repair deficiency and its impact on tumour characteristics

(CIN in this case). To this end, we calculated for each breast tumour an *HR* score that quantified the extent of HR pathway dysregulation in that tumour. Based on the score, we observed a great diversity in HR dysregulation between and within the PAM50 subtypes, and by using two published HR-defect signatures, we found HR pathway dysregulation reflects HR repair deficiency. More importantly, we uncovered a novel association between HR dysregulation and CIN, which indicates that dysregulated HR might contribute to replication stress-induced CIN in breast cancer. This knowledge may help future studies to identify the causative factors of CIN in sporadic breast cancer as well as in other cancer types.

Materials and Methods

1. Genomic data

Whole-genome gene expression data, DNA copy-number data, gene mutation data (only available for the TCGA samples) and related clinical data for four breast cancer cohorts (Table 1) were obtained from METABRIC (Curtis et al., 2012) and TCGA (Koboldt et al., 2012).

Table 1 Breast cancer cohorts analysed in this study. The METABRIC gene expression data are microarray-based.

Cohort	No. of tumour samples						No. of normal breast tissues
	All	Basal-like	HER2	LumA	LumB	Normal-like	
METABRIC Discovery	997	118	87	466	268	58	144
METABRIC Validation	995	213	153	255	224	144	144
TCGA RNA-seq	1068	188	80	549	213	38	113
TCGA Microarray	522	98	58	231	127	8	22

Gene-expression data and chromosomal-level DNA copy-number data from the METABRIC project (Genome-phenome Archive accession number EGAS00000000083) were made available upon request, and had already been preprocessed as described by Curtis et al. (Curtis et al., 2012). Gene-expression data from this project were based on the Illumina HT-12 v3 Expression Beadchip (Illumina, San Diego, CA, USA). The probe-level transcription estimates were mapped to gene-level estimates using the HT-12 v3 annotation file downloaded from the Illumina website (http://www.illumina.com/). Where two or more probes represented the same gene, the probe with the largest variation was chosen as the gene representative. DNA copy-number data from METABRIC had been generated using Affymetrix SNP 6.0 arrays (Affymetrix, Santa Clara, CA, USA). The corresponding PAM50 subtype assignment and clinical outcome were obtained from (Curtis et al., 2012).

The preprocessed gene-expression and DNA copy-number data (both chromosome-level and gene-level) for the TCGA RNA-seq cohort were downloaded via the UCSC Cancer Genomics

Browser (https://genome-cancer.ucsc.edu/) on 13 October 2014. Gene-expression data for this cohort were measured using the Illumina HiSeq 2000 RNA Sequencing platform, and show the Expectation Maximization (RSEM)-normalised and percentile-ranked gene-level transcription estimates. DNA copy-number data for this cohort had been generated using Affymetrix SNA 6.0 arrays, with germline copy-number variation filtered out. PAM50 classifications for this cohort were obtained through personal communication with the TCGA consortium. A subset of these 1068 cases also has gene expression data obtained from microarray. The Level 3 gene-expression data for this TCGA Microarray cohort and the corresponding PAM50 classifications were downloaded from the TCGA data portal publication site (https://tcga-data.nci.nih.gov/docs/publications/brca_2012/) on 3 June 2014. These gene-expression data were based on Agilent custom 244K whole-genome microarrays and had been preprocessed as described by Koboldt et al. (Koboldt et al., 2012). DNA copy-number data for this cohort were obtained as a subset of the TCGA RNA-seq cohort, as the samples of the former cohort were covered by the later cohort.

The preprocessed gene mutation data for 982 TCGA samples, generated on an IlluminaGA system, were downloaded via the UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/) on 6 July 2015. Each gene had been assigned a value of 1 or 0, indicating whether a non-silent mutation was identified in the coding region of that gene (value=1) or not (value=0). These data were matched to the two TCGA cohorts respectively according to the sample ID.

2. HR pathway curation and calculation of HR score

Based on our recent review of the HR pathway (Liu et al., 2014), we manually curated a list of 82 genes with direct relevance to HR (Supplementary Table S1). We then applied Pathifier (Drier et al., 2013) to the mRNA expression level of the curated HR genes to calculate an *HR* score that quantifies HR pathway dysregulation in individual breast tumours. Based on gene-expression profiles for tumours and normal breast tissues, Pathifier transforms HR gene-expression measurements into a measure of HR pathway dysregulation by fitting a principal curve (see Supplementary Figure S1 for a visualisation of the curve) that captures the maximal variability of the expression levels of the HR genes in all samples, and then projects each sample onto that curve. A sample's *HR* score is defined as its distance along the curve from the centroid of the normal tissues (Drier et al., 2013).

Not all HR genes we curated were present in the gene expression data for each of the four cohorts. We therefore calculated the *HR* score for each cohort based only on HR genes that are available for that cohort (ranges from 67 to 72, see Supplementary Table S1). No other ways for selecting HR genes were examined to minimize retrospective optimization for the correlations with CIN (see below).

3. CIN measurements calculation

The numbers of chromosomal breakpoints and the proportions of the genome affected by copy-number change (Genomic Instability Index, GII) for samples in the two METABRIC cohorts were downloaded from a recent study (Vollan et al., 2015) in which the METABRIC Group was involved. According to this study, a few samples with mismatched DNA/RNA were identified and excluded, resulting in 985 samples remaining in the Discovery cohort and 965 in the Validation cohort. To get the number of amplified/deleted genes for the same samples, we first calculated the copy number of each gene using the chromosomal-level DNA copy-number data available for the two cohorts, then applied cut-offs (≥ 0.10 for amplified genes and ≤ -0.15 for deleted genes; values represent log2 ratios of the tumour intensity to the normal intensity) that are similar to those used by METABRIC to define chromosomal regions with amplifications or deletions.

For the two TCGA cohorts, we used the chromosomal-level DNA copy-number data to calculate number of breaks by counting the total number of chromosomal segments at least 1 kb in length. The calculation of GII was also based on the chromosomal-level DNA copy-number data after filtering out segments shorter than 1kb, and the same cut-offs as mentioned above (≥ 0.10 for amplification and ≤ -0.15 for deletion) were used to identify chromosomal regions with copy-number change. The number of amplified/deleted genes for each of the two TCGA cohorts was obtained from the downloaded gene-level DNA copy-number data, where +1 and +2 represent amplification and -1 and -2 represent deletion.

4. Survival analysis

Survival analysis for both of the METABRIC datasets was performed using the R package *survival* (http://cran.r-project.org/web/packages/survival/index.html). Patient follow-up time was limited to 15 years, and only breast cancer-related deaths were counted.

Results

1. An HR score for quantifying HR pathway dysregulation in individual breast tumours

An *HR* score was developed for each breast tumour to quantify HR pathway dysregulation in that tumour; a high *HR* score means that the expression of the HR genes as a whole in an individual tumour is very different from the situation in normal breast tissues (see Supplementary Figure S2 for HR gene expression in tumours with low to high *HR* score). To calculate this score, we first manually curated a list of 82 HR genes (Supplementary Table S1) according to our recent review on the HR pathway (Liu et al., 2014). This gene list provides more up-to-date knowledge

about the content of HR compared to publicly available pathway databases; for instance, it catalogues 54 more genes than the HR pathway in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000). The expression profiles of the curated HR genes were then employed as input to the Pathifier method (Drier et al., 2013) to compute the score. To ensure reproducibility of the results, we performed this pathway analysis independently for four large breast cancer cohorts that also include data on normal breast tissues (Table 1). Depending on data availability, the number of HR genes for calculating the score is slightly different across the cohorts (Supplementary Table S1).

The boxplots in Figure 1 display the *HR* score distribution in each cohort with regard to the PAM50 molecular subtypes, and in normal breast tissues. We observed a consistent pattern across the four cohorts: basal-like tumours generally have the highest *HR* score, followed by HER2 and Luminal B tumours, and then Luminal A and Normal-like tumours; the normal breast tissues always have the lowest *HR* score as a consequence of being the benchmark. Similar results can be seen in Supplementary Figure S3 showing *HR* score versus the *HR* score-based rank of the tumours of different subtypes. The consistent distribution of the *HR* score by tumour subtype across the different cohorts and gene-expression profiling platforms (RNA-seq and microarray in TCGA) is strong evidence that the *HR* score is robust and reproducible. Interestingly, we observed some variability in *HR* score within tumours of the same subtype, as highlighted by some outliers in the boxplots, suggesting some heterogeneity in HR pathway dysregulation within the subtypes.

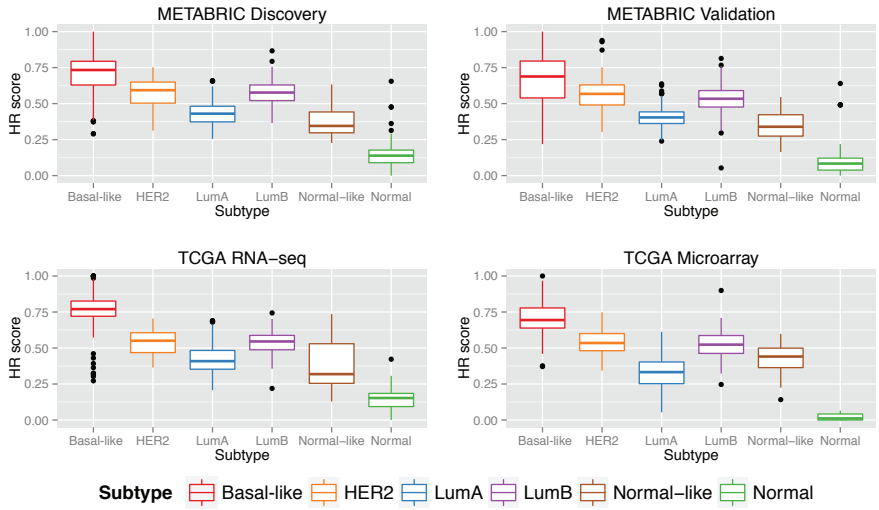


Figure 1 Distribution of the *HR* score across the PAM50 subtypes and normal breast tissues (in green) for the four cohorts. The scores were generated independently within each cohort.

2. The *HR* score is reflective of HR repair deficiency.

The *HR* score is gene expression-based, and measures the extent to which the HR pathway is dysregulated. To test whether there exists an association between HR pathway dysregulation and

HR repair deficiency, we next asked whether the *HR* score is reflective of HR repair deficiency (i.e., whether a tumour with high *HR* score is likely to be HR-defective). We used two published HR-defect signatures, homologous recombination defect (HRD) (G. Peng et al., 2014) and large-scale transitions (LSTs) (Popova et al., 2012), to test this hypothesis.

2.1. Comparison with the HRD signature

The HRD signature encompasses 230 genes that are differentially expressed between HR-intact and HR-deficient cells, and is intended to represent the global impact of HR defect on the transcriptome of a tumour cell (G. Peng et al., 2014). To identify tumours (or cell lines) with HR deficiency, Peng et al. performed a hierarchical clustering analysis based on the expression level of the 230 genes to divide samples into two clusters, one considered as HR-intact and the other HR-deficient (G. Peng et al., 2014).

In this study, we performed the same clustering analysis for each of the four cohorts (Figure 2A for the METABRIC Discovery cohort and Supplementary Figures S4, S5 and S6 for the three remaining cohorts). As shown in Figure 2A, tumours with low HR score (upper horizontal bar, green) are mostly tumours belonging to the HR-intact cluster, whereas tumours with high HR score (upper horizontal bar, red) are mostly tumours belonging to the HR-deficient cluster. To be more precise, Figure 2B shows the distribution of the HR score in the two HRD-based clusters for each of the four cohorts, demonstrating that tumours in the HR-deficient cluster in general have significantly higher HR score compared with tumours in the HR-intact cluster (p-values $\leq$ 9.1e-63, Wilcoxon rank-sum test). These observations indicate that tumours with high HR scores are likely to be HR-defective, as predicted by the HRD signature.

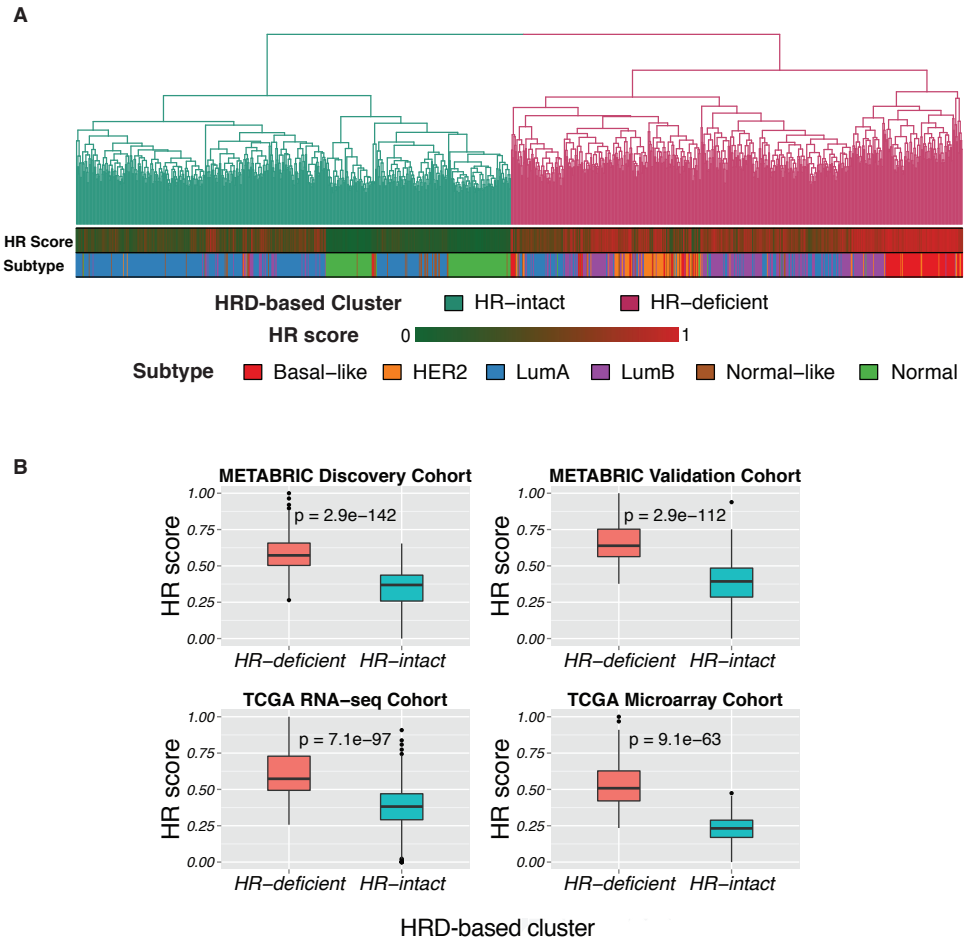


Figure 2 Comparison of the *HR* score with the HRD signature. A: HRD-based hierarchical clustering of tumours from the METABRIC Discovery cohort. B: Distribution of the *HR* score in the two HRD-based clusters for each of the four cohorts. Colour represents the HRD-based cluster. The p-values were obtained using a Wilcoxon rank-sum test.

2.2. Comparison with the LST signature

LST refers to a chromosomal break whose flanking regions are at least 10 Mb in size. A tumour with a large number of LSTs indicates HR defect-related genomic scarring as a measure of chromosomal instability (Popova et al., 2012). In this study, we estimated the number of LSTs for each tumour using the DNA copy number data, and divided each cohort into two groups according to the method and cut-offs described in (Popova et al., 2012): LST^+ (≥ 20 LSTs) and LST^- (< 20 LSTs). The numbers of LST^+ and LST^- tumours identified in each cohort are summarised in Supplementary Table S2. As in the comparison with the HRD signature, we found that LST^+ tumours generally have higher HR scores compared with LST^- tumours, even in the case of the METABRIC Discovery cohort where only nine LST^+ tumours were identified (Figure 3). This observation also supports the idea that the *HR* score is indicative of HR defect.

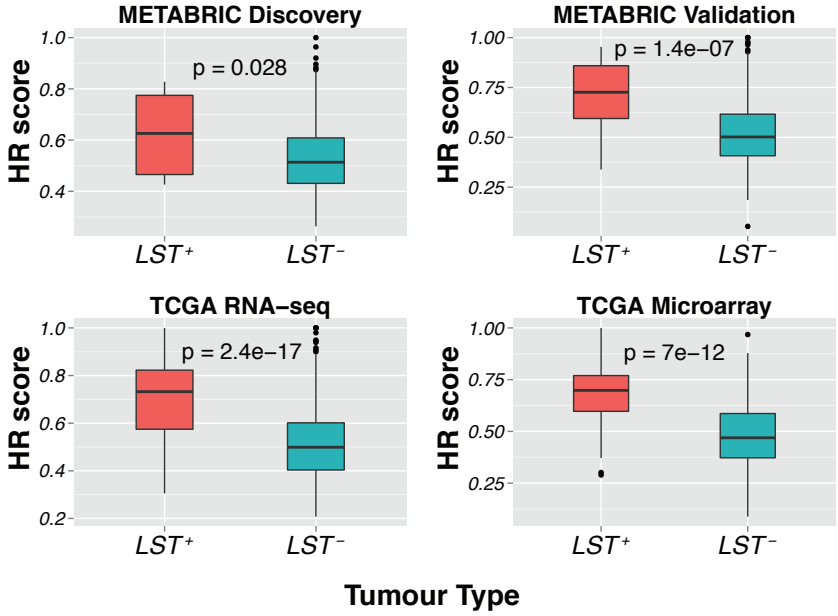


Figure 3 Distribution of the *HR* score in LST$^+$ tumours and LST$^-$ tumours for each of the four cohorts. Colour represents LST status. The p-values were obtained using a Wilcoxon rank-sum test.

Taken together, the results based on HRD and on LST demonstrate an association between HR pathway dysregulation, as represented by the *HR* score, and HR repair deficiency. In addition, in the two TCGA cohorts for which gene mutation data are available, we also observe that tumours with at least one non-synonymous mutation in one of six key HR genes have significantly higher *HR* score than do the tumours with no mutation in any of these genes (see Supplementary Figure S7 for more details). All these results support the existence of a compensatory mechanism through which HR-deficient cells respond to their HR defect by altering the expression level of HR genes. Interestingly, it has been proposed that melanoma cells exploit the overexpression of DNA repair genes, particular those involved in DSB repair, to increase their DNA repair capacity that is necessary for them to invade and give rise to distant metastases (Sarasin and Kauffmann, 2008). Consistent with this, overexpression of certain DNA repair genes is utilised by polyploid cells to overcome replication stress-induced senescence barriers (Zheng et al., 2012). All these results indicate that altering the expression of DNA repair genes or pathways may be a compensatory mechanism commonly exploited by tumour cells.

3. Association with CIN

Because replication stress has emerged as a common source of CIN in caner, and HR is the crucial pathway for the repair of replication stress-induced DSBs, we hypothesised that there might be a link between HR pathway dysregulation, which is indicative of HR repair deficiency as described above, and the degree of CIN in breast carcinomas. To test this hypothesis, we first examined the correlation between the *HR* score and the widely used CIN signature CIN25 (Carter et

al., 2006). We then investigated the association between the *HR* score and each of the three common CIN measurements: number of chromosomal breakpoints, fraction of the genome with copy-number alterations (genomic instability index, GII), and number of amplified/deleted genes. In particular, as data pre-processing and segregation algorithms can significantly affect the actual value of the CIN measurements, we downloaded the numbers of chromosomal breaks and GII for the two METABRIC cohorts from a recent publication (Vollan et al., 2015). We believe these measures from a third-party study provide more-objective results for our analysis.

3.1. Association with CIN25

Figure 4 displays a scatter plot between the CIN25 score, defined as the mean expression value of the CIN25 genes (Carter et al., 2006), and the *HR* score for tumours from each of the four cohorts. Each cohort showed a high correlation between the CIN25 score and the *HR* score (Spearman correlation coefficient r = 0.94 and r = 0.93 for the two METABRIC cohorts, and r = 0.85 and r = 0.96 for the two TCGA cohorts), indicating that the *HR* score is also correlated with CIN level. Moreover we found ten of the CIN25 genes (40%) to be present among the 230 genes of the HRD signature mentioned in Section 2.1, which indicates that HR defects might be one of the underlying biological mechanisms responsible for the expression change of the CIN25 genes.

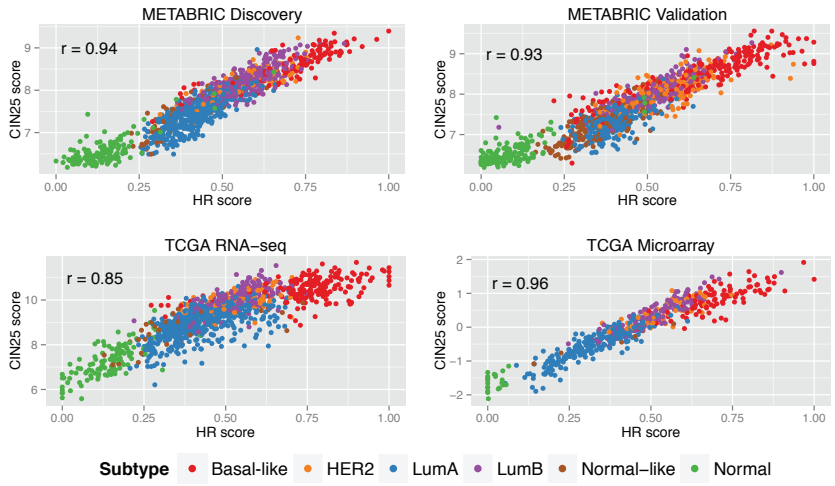


Figure 4 Correlations between the CIN25 score and the *HR* score for each of the four cohorts.

Overall, these results revealed that the *HR* score correlates with the CIN25 score, and support the hypothesis that there exists an association between HR pathway dysregulation, as represented by the *HR* score, and CIN level in tumours, as predicted by the CIN25 score.

3.2. Association with three common CIN measurements

Because the CIN25 score only indirectly estimates CIN level in tumours, we also directly assessed the relationship between the *HR* score and each of the three common CIN measures (breakpoints, GII and number of amplified/deleted genes). We asked whether tumours with higher

HR score tend to have a higher CIN level. To address this, we divided tumours into four equal-sized groups based on the *HR* score quartiles, and statistically examined the differences between adjacent groups for each of the three CIN measurements. The boxplots in Figure 5 (METABRIC Discovery cohort) show a high variability in each *HR* score quartile group for each CIN measurement, indicating that other mechanisms can also affect CIN. However, we observed a clear pattern that tumours with higher *HR* score indeed tend to have higher CIN level (Wilcoxon rank-sum test, one sided FDR p-value < 0.05), with the exception of tumours in the third and fourth quartile groups in GII. Similar results were obtained for the remaining three cohorts (Supplementary Figures S8, S9 and S10). Overall, these results suggest an association between the extent of HR pathway dysregulation and the degree of CIN level in breast carcinomas.

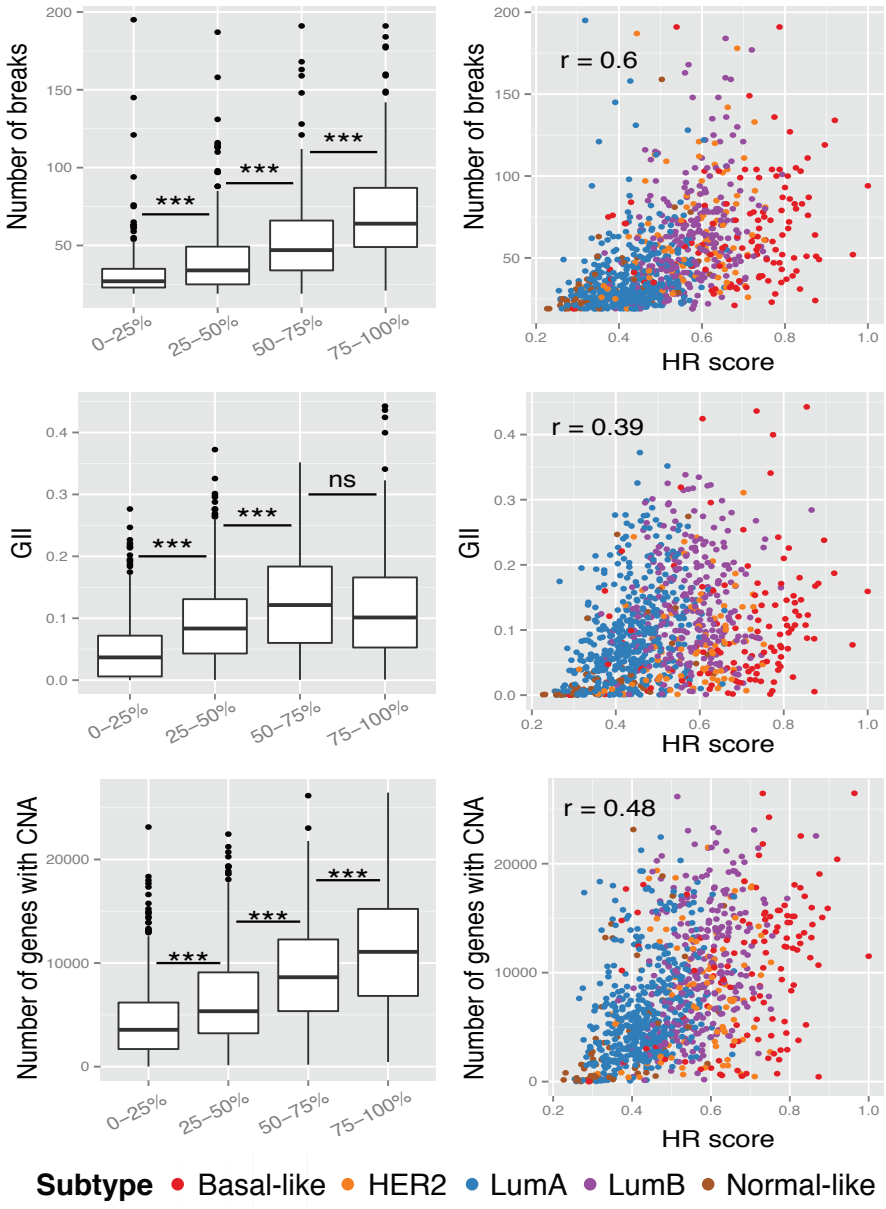


Figure 5 *HR* score versus the three CIN measurements for the METABRIC Discovery cohort. Left: Boxplots of the three CIN measurements versus the four *HR* score quartile groups; stars indicate statistical significance according to a Wilcoxon rank-sum test: ns means not significant and *** means p-value < 0.001. Right: Scatter plots of the *HR* score versus each of the three CIN measurements; r represents Pearson Correlation Coefficient.

As the *HR* score is based on gene expression, to ascertain whether the association observed above is due to the gene expression-based PAM50 subtypes, we performed the same analysis independently on tumours within each PAM50 subtype. In each analysis, the samples were divided into high and low *HR* score groups according to the median. The results for the METABRIC Discovery cohort are summarised in Figure 6. For this cohort we consistently observed that tumours in the high *HR* score group have more breakpoints than do tumours in the low *HR* score group within the subtypes, despite the wide range of the breakpoint numbers observed for each subtype. The difference in GII between the low and high *HR* score groups was significant in Basal-like, Luminal A and Normal-like tumours, but not in HER2 and Luminal B tumours, while the difference in number of amplified/deleted genes between the two groups was significant in all subtypes except HER2. For the other cohorts (Supplementary Figures S11, S12 and S13) we observed some differences between cohorts. For example, in the METABRIC Validation cohort, all three CIN measurements are significantly different between the two *HR* score groups for all subtypes, whereas the difference is significant in fewer subtypes in the TCGA Microarray cohort. These discrepancies might be due to low sample size in the TCGA Microarray cohort (e.g. there are only eight samples in its Normal-like subtype). Apart from these possible exceptions, the above results support the hypothesis that tumours with more-deregulated HR pathway are likely to have a higher degree of CIN, and this relationship can still be detected within the gene expression-based PAM50 subtypes.

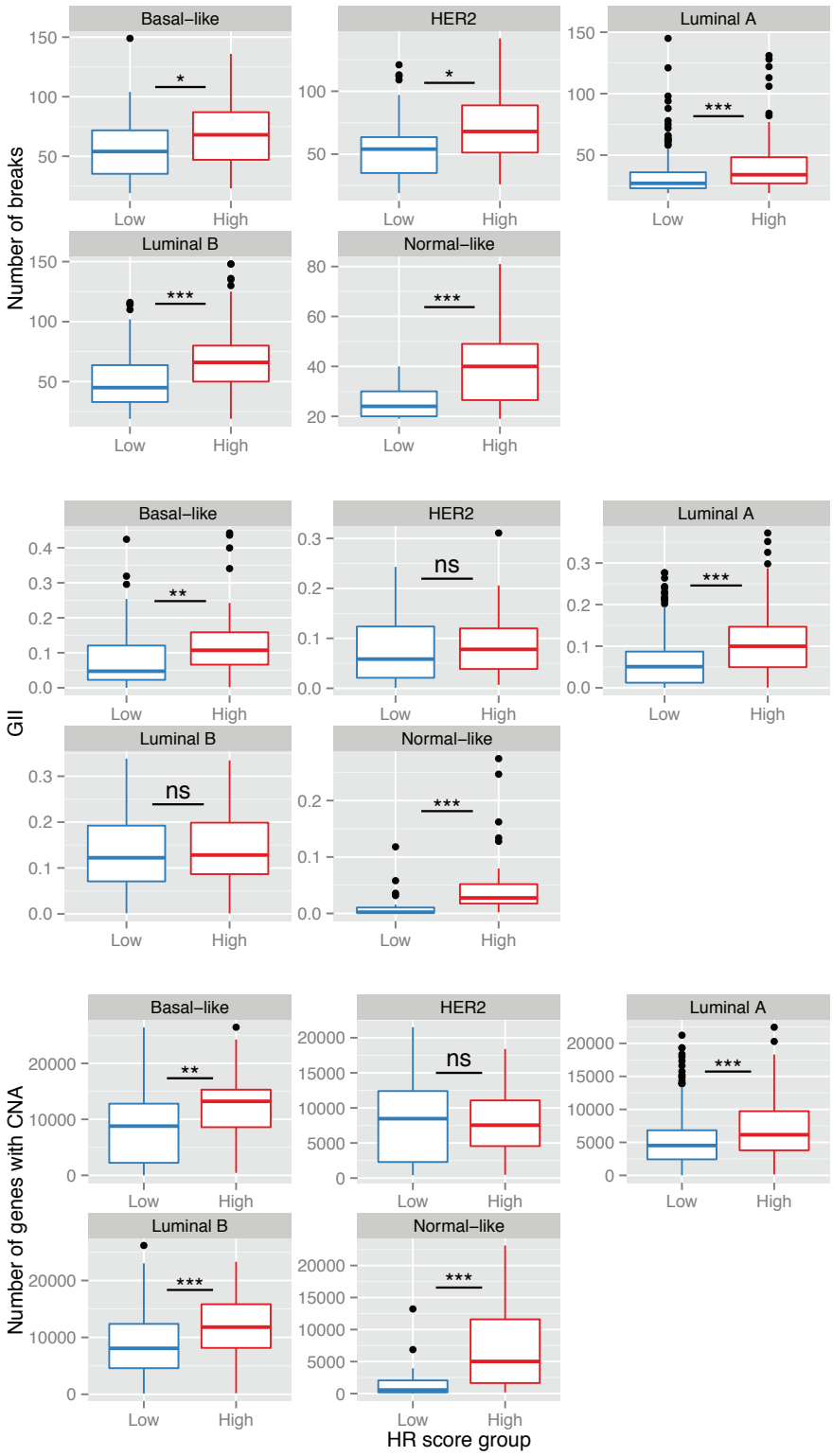


Figure 6 *HR* score versus the three CIN measurements within PAM50 subtypes (METABRIC Discovery cohort). For each plot, the two *HR* score groups were divided according to the median *HR* score in each subtype; stars indicate the significance according to a Wilcoxon rank-sum test for each pair of groups: ns means not significant, * means $0.01< p < 0.05$, ** means $0.001 < p <0.01$, and *** means $p < 0.001$.

3.3. Association between the CIN measurements and other pathways

The scatter plots in Figure 5 (METABRIC Discovery cohort) show that the *HR* score is moderately correlated with each of the three CIN measurements (breakpoints r = 0.60, GII r=0.39

and number of amplified/deleted genes r = 0.48). These moderate correlations are not surprising, given that we do not consider aberrant HR as the only mechanism that contributes to CIN. In this section we investigated whether there are other pathways whose dysregulation also correlates with CIN, and whether these moderate correlations are far from random.

We computed a score for each of the 186 KEGG pathways (Kanehisa and Goto, 2000) and for 674 Reactome pathways (Croft et al., 2010), using the same approach as for the *HR* score. Spearman correlation coefficients between these scores and each of the three CIN measures were recorded and compared against the respective correlations between the *HR* score and the three CIN measurements. Figure 7 shows the results for the METABRIC Discovery cohort (KEGG pathways are in green and Reactome pathways in blue; similar results for the other three cohorts are in Supplementary Figures S14, S15 and S16). We found only a few KEGG or Reactome pathways whose dysregulation showed a similar level of correlation with CIN as did the HR pathway. For example, only four (2.2%) KEGG pathways (cell cycle, oocyte meiosis, progesterone-mediated oocyte maturation and p53 signalling) were more strongly associated with number of breakpoints than with the HR pathway (r = 0.61 - 0.63 compared to r = 0.60 for the HR pathway in Figure 7). Moreover, the strong associations of the oocyte meiosis, progesterone-mediated oocyte maturation and p53 signalling pathways with number of breakpoints is mainly due to their considerable overlap in gene content with the KEGG cell cycle pathway: 37%, 34% and 36% genes from each of these three pathways are also present in the cell cycle pathway (Supplementary Table S3). In contrast, only two HR genes are present in the cell cycle pathway. After removing the overlapping genes, association levels between each of these three pathways with number of breakpoints significantly decreased (results not shown). Similarly, although there were 24 (3.6%) Reactome pathways whose dysregulation showed a similar level of correlation with CIN as did the HR pathway, 18 of these are either the cell cycle pathway or its sub-pathways (Supplementary Table S4).

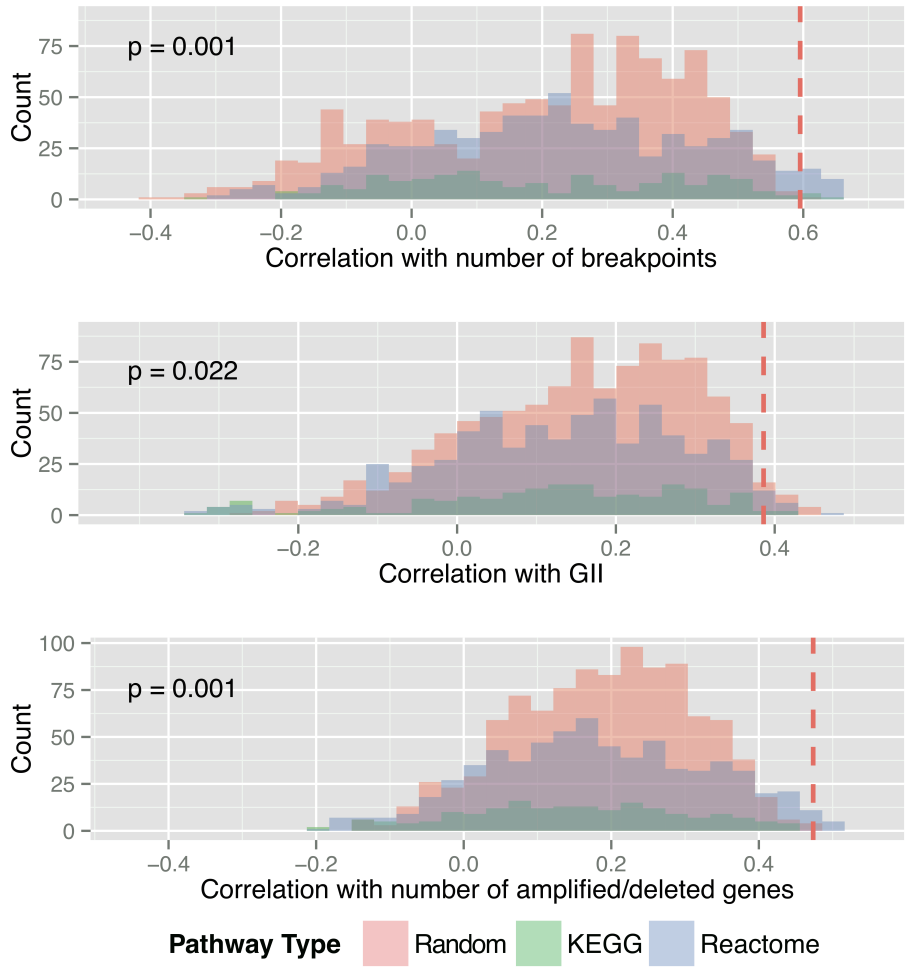


Figure 7 Distributions of the correlations between pathway scores and the three CIN measurements (METABRIC Discovery cohort). Results for KEGG pathways are in green, Reactome pathways in blue and Random pathways in pink. Spearman correlation coefficients (r) are represented on the x-axis. Pathway score were calculated with Pathifier. The vertical dashed line in each histogram indicates the value of r between the *HR* score and each of the three CIN measurements, and p represents an empirical p-value for that value of r.

As the KEGG and Reactome pathways do not cover all genes measured in the whole-genome gene expression profiling data analysed in this study, we also constructed 1000 "Random" pathways for each cohort to calculate an empirical p-value for the association between the HR score and each of the three CIN measurements. Each Random pathway is of the same length as HR but is composed of genes randomly selected from the gene-expression profiling data, excluding those from HR and cell cycle pathways. Similar to the KEGG pathways analysed above, we computed a score for each Random pathway, and compared the correlation coefficients with the three CIN measures against those for the *HR* score. As shown in Figure 7, only a few Random pathways (in pink) showed a level of association with CIN similar to that of the HR pathway, as indicated by the empirical p-values. Similar results for the other three cohorts were obtained (Supplementary Figures S14, S15 and S16).

Overall, these results indicate that the CIN level in tumours is associated with the dysregulation of only a limited number of pathways (e.g., the cell cycle pathway), and that the correlation between HR and CIN is far from being random.

4. Association with survival in ER$^+$ tumours

The two METABRIC cohorts are annotated with disease-specific survival data that are lacking for the two TCGA cohorts. We thus tested whether the *HR* score can predict patient survival in the two METABRIC cohorts. Figure 8 shows Kaplan-Maier plots for patients with ER$^+$ tumours from the METABRIC Discovery (n=699; follow-up time ≤ 15 years) and validation cohorts (n=582; follow-up time ≤ 15 years). For each cohort, patients were divided into high and low *HR* score groups based on the median *HR* score. For both cohorts, we observed a significant difference in patient survival between the two *HR* score groups with ER$^+$ tumours (Figure 8; Cox proportional hazards regression test p-value = 8.4e-04 and 3.9e-09 for the two cohorts, respectively). However, we observed no significant difference in survival between the two *HR* score groups for patients with ER$^-$ tumours (data not shown). As an association between CIN and prognosis in ER$^+$ tumours has already been documented (Przybytkowski et al., 2014; Smid et al., 2011), and after control for the number of chromosomal breaks there is no significant difference in survival between the two *HR* score-based groups (result not shown), we infer that the prognostic value of the *HR* score in ER$^+$ tumours is due to the association between the *HR* score and CIN.

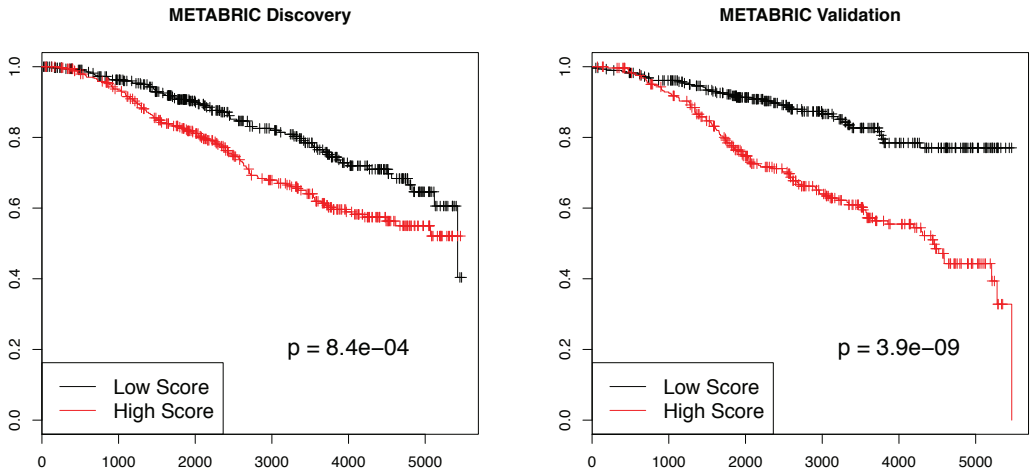


Figure 8 Kaplan-Maier plot for disease specific survival in the METABRIC Discovery cohort (left) and Validation cohort (right). Patients with ER$^+$ tumour were divided into two equal-sized groups based on the median *HR* score in each cohort.

Discussion

Multiple molecular mechanisms have been associated with the origin of CIN in cancer, including replication stress, telomere dysfunction, aberrant DNA repair and various defects in chromosome segregation (reviewed in Abbas et al., 2013; Aguilera and García-Muse, 2013; Negrini et al., 2010; Thompson et al., 2010). Although CIN can be experimentally induced by exploiting any of these mechanisms, replication stress has been recently identified as the first recurrent genetic defect associated with CIN in colorectal cancer (Burrell et al., 2013a). In this scenario, CIN is induced during DNA replication in fast-dividing tumour cells, giving rise to frequent stalling of replication forks. Consequently, HR as the primary pathway for repair of the resultant DSBs during replication becomes overworked, and if HR is dysfunctional the frequency of replication stress-induced CIN is likely to increase dramatically. Here we have shown that HR dysregulation as measured by the *HR* score, which is indicative of aberrant HR repair, is prevalent in sporadic breast cancer and correlates with the level of CIN. We thus propose that HR dysregulation might contribute to replication stress-induced CIN at least in sporadic breast cancer. Consistent with this view, overexpression of the key HR gene *RAD51*, which is commonly seen in breast cancer as well as other cancer types, promotes chromosomal instability (Richardson et al., 2004), and two other critical HR genes, *BRCA1* and *BRCA2*, were recently proposed as chromosome custodians mainly due to their role in HR (Venkitaraman, 2014a; 2014b).

Dysfunction of the HR pathway, although not the primary cause, may increase the level of replication stress-induced CIN in several ways. Firstly, it can cause inefficient repair of DSBs, resulting in an accumulation of chromosomal breaks. Secondly, by triggering error-prone repair pathways including canonical non-homologous end-joining (C-NHEJ) and alternative non-homologous end-joining (Alt-NHEJ, also called microhomology-mediated end joining (MMEJ)), HR dysfunction can lead to translocations, translocation-related chromosomal breaks and DNA copy-number changes. Specifically, in contrast to HR that requires homologous sequence to guide repair, C-NHEJ and Alt-NHEJ mediate the repair by a direct ligation of the break ends after more-or-less end processing, and so do not ensure that the broken DNA strands are re-joined in the correct position. These two low-fidelity pathways come to repair DSBs generated during DNA replication when HR is deficient, resulting in translocation as well as translocation-related chromosomal breaks (Alexandrov et al., 2013; Bunting and Nussenzweig, 2013; Ottaviani et al., 2014; Villarreal et al., 2012). Moreover, gene copy number changes also arise when the repair of broken replication forks switched from HR to the two NHEJs, especially Alt-NHEJ (Hastings et al., 2009);

A third way in which HR pathway dysfunction can increase replication stress-induced CIN is by affecting mitosis and the proper functioning of telomeres. HR defects and the consequent slow progression of replication forks can elicit alterations of mitosis, which highlights the importance of HR at the interface of these two processes for protection against CIN (Wilhelm et al., 2014). In addition, DSB repair is shut down during the M phase to avoid telomere fusion and as a consequence, mitosis will continue even in the presence of DSBs or fragmented chromosomes, giving rise to CIN (Orthwein et al., 2014). This emphasises the importance of DSB repair during DNA replication, especially given the presence of DSBs that result from replication stress. HR defects caused by *BRCA2* mutations could also lead to telomere dysfunction, a mechanism that has been proposed to explain, in part, the chromosomal instability observed in *BRCA2*-deficient tumours (Badie et al., 2010). Taken together, HR dysfunction can increase CIN via diverse mechanisms, and the association revealed in this study between HR dysregulation and CIN (Figures 4, 5 and 6) indicates that dysregulated HR might contribute to the CIN observed in highly replicative tumours.

The study of CIN in breast cancer has attracted immense interest in recent years following the recognition of its clinical relevance in disease heterogeneity, drug resistance and patient response (A'Hern et al., 2013; Birkbak et al., 2011; Endesfelder et al., 2014; Habermann et al., 2009; Roylance et al., 2011; Sansregret and Nepveu, 2011; Swanton et al., 2009; Vincent-Salomon et al., 2013); reviewed by (Wiechec, 2011). CIN induces evolution in tumours, providing the heterogeneity from which aggressive and/or drug-resistant tumour clones are selectively established. CIN aids tumour development by amplifying genomic regions containing oncogenes and deleting regions containing tumour-suppressor genes, thereby significantly influencing treatment response and survival in patients. Our results further strengthen this connection by associating dysregulated HR with the extent of amplified/deleted genes and regions of the chromosome, and by showing that ER^+ tumours with high HR score or CIN levels display significantly poorer prognosis (Figure 8).

A measure of HR dysregulation such as the one adopted here can be extremely valuable to guide therapeutic options. The observation that cancer cells deficient in HR are profoundly sensitive to PARP inhibitors (Bryant et al., 2005; Farmer et al., 2005) has already led to the development of targeted PARP therapies for sporadic breast and ovarian cancers with defects in core HR genes such as *BRCA1* and *BRCA2*, a condition termed as "BRCAness" (Turner et al., 2004). PARP is an important protein family whose members function in restarting stalled replication forks and diverting DSBs to HR-mediated repair. It has been proposed that accumulated chromosomal instability arising from the continued stalling of replication forks, accompanied by deficiency in repairing DSBs and thereby triggering a genomic catastrophe, may explain how PARP inhibition

kills HR-deficient cancer cells (Bryant et al., 2005; Farmer et al., 2005). Although focussing on a mechanistic explanation for PARP-based cancer therapy, these models indirectly suggest an underlying relationship among replicative stress, dysfunctional HR and the accumulation of chromosomal instability.

In conclusion, we performed a personalised pathway analysis by calculating an *HR* score that quantifies HR pathway dysregulation in individual breast tumours, with the behaviour of HR in normal breast tissues serving as a benchmark. Our results are reproducible across four large breast cancer cohorts (~ 3000 tumours in total). We found HR is dysregulated to various extents between and within the gene expression-based PAM50 subtypes, which may reflect their HR repair deficiency. More importantly, we uncovered a novel association between HR dysregulation and CIN. Although HR has a well-known role in maintaining genomic integrity, this work is the first large-scale study to assess the correlation between HR dysregulation and CIN in sporadic breast cancer. As such our results will be useful for future studies that aim to identify causative factors of CIN in sporadic breast cancer as well as in other cancer types.

Acknowledgements

This study makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium funded by Cancer Research UK and the British Columbia Cancer Agency Branch. We also thank TCGA for providing the genomic data. This study was funded by the Australian National Health and Medical Research Council (NHMRC) Project Grant (ID: 1028742) to PTS and MAR. KALC was supported in part by the Australian Cancer Research Foundation (ACRF) for the Diamantina Individualised Oncology Care Centre at The University of Queensland Diamantina Institute and the NHMRC Career Development fellowship (ID: 1087415). KKK is an NHMRC Senior Principal Search Fellow (ID: 613638) supported by the NHMRC Project Grant (ID: 1017028).

References

A'Hern, R.P., Jamal-Hanjani, M., Szász, A.M., Johnston, S.R.D., Reis-Filho, J.S., Roylance, R., Swanton, C., 2013. Taxane benefit in breast cancer--a role for grade and chromosomal stability. Nat Rev Clin Oncol 10, 357–364. doi:10.1038/nrclinonc.2013.67

Abbas, T., Keaton, M.A., Dutta, A., 2013. Genomic instability in cancer. Cold Spring Harbor Perspectives in Biology 5, a012914. doi:10.1101/cshperspect.a012914

Aguilera, A., García-Muse, T., 2013. Causes of genome instability. Annu. Rev. Genet. 47, 1–32. doi:10.1146/annurev-genet-111212-133232

Ahn, T., Lee, E., Huh, N., Park, T., 2014. Personalized identification of altered pathways in cancer using accumulated normal tissue data. Bioinformatics 30, i422–9. doi:10.1093/bioinformatics/btu449

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V.,

Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al., 2013. Signatures of mutational processes in human cancer. Nature 500, 415–421. doi:10.1038/nature12477
Badie, S., Escandell, J.M., Bouwman, P., Carlos, A.R., Thanasoula, M., Gallardo, M.M., Suram, A., Jaco, I., Benitez, J., Herbig, U., Blasco, M.A., Jonkers, J., Tarsounas, M., 2010. BRCA2 acts as a RAD51 loader to facilitate telomere replication and capping. Nature Structural & Molecular Biology 17, 1461–1469. doi:10.1038/nsmb.1943
Bekker-Jensen, S., Mailand, N., 2011. The ubiquitin- and SUMO-dependent signaling response to DNA double-strand breaks. FEBS Letters 585, 2914–2919. doi:10.1016/j.febslet.2011.05.056
Birkbak, N.J., Eklund, A.C., Li, Q., McClelland, S.E., Endesfelder, D., Tan, P., Tan, I.B., Richardson, A.L., Szallasi, Z., Swanton, C., 2011. Paradoxical relationship between chromosomal instability and survival outcome in cancer. Cancer Research 71, 3447–3452. doi:10.1158/0008-5472.CAN-10-3667
Brown, E.T., Holt, J.T., 2009. Rad51 overexpression rescues radiation resistance in BRCA2-defective cancer cells. Mol. Carcinog. 48, 105–109. doi:10.1002/mc.20463
Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J., Helleday, T., 2005. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. Nature 434, 913–917. doi:10.1038/nature03443
Bunting, S.F., Nussenzweig, A., 2013. End-joining, translocations and cancer. Nat Rev Cancer 13, 443–454. doi:10.1038/nrc3537
Burrell, R.A., McClelland, S.E., Endesfelder, D., Groth, P., Weller, M.-C., Shaikh, N., Domingo, E., Kanu, N., Dewhurst, S.M., Gronroos, E., Chew, S.K., Rowan, A.J., Schenk, A., Sheffer, M., Howell, M., Kschischo, M., Behrens, A., Helleday, T., Bartek, J., Tomlinson, I.P., Swanton, C., 2013a. Replication stress links structural and numerical cancer chromosomal instability. Nature 494, 492–496. doi:10.1038/nature11935
Burrell, R.A., McGranahan, N., Bartek, J., Swanton, C., 2013b. The causes and consequences of genetic heterogeneity in cancer evolution. Nature 501, 338–345. doi:10.1038/nature12625
Carter, S.L., Eklund, A.C., Kohane, I.S., Harris, L.N., Szallasi, Z., 2006. A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers. Nat. Genet. 38, 1043–1048. doi:10.1038/ng1861
Chowdhury, D., Choi, Y.-E., Brault, M.E., 2013. Charity begins at home: non-coding RNA functions in DNA repair. Nat Rev Mol Cell Bio 14, 181–189. doi:10.1038/nrm3523
Ciriello, G., Sinha, R., Hoadley, K.A., Jacobsen, A.S., Reva, B., Perou, C.M., Sander, C., Schultz, N., 2013. The molecular diversity of Luminal A breast tumors. Breast Cancer Res Treat 141, 409–420. doi:10.1007/s10549-013-2699-3
Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., Jupe, S., Kalatskaya, I., Mahajan, S., May, B., Ndegwa, N., Schmidt, E., Shamovsky, V., Yung, C., Birney, E., Hermjakob, H., D'Eustachio, P., Stein, L., 2010. Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Research 39, D691–D697. doi:10.1093/nar/gkq1018
Curtis, C., Shah, S.P., Chin, S.-F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., Gräf, S., et al., 2012. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 1–7. doi:10.1038/nature10983
d'Adda di Fagagna, F., 2014. A direct role for small non-coding RNAs in DNA damage response. Trends in Cell Biology 24, 171–178. doi:10.1016/j.tcb.2013.09.008
Davis, J.D., 2011. DNA damage and breast cancer. WJCO 2, 329. doi:10.5306/wjco.v2.i9.329
Dou, H., Huang, C., Van Nguyen, T., Lu, L.-S., Yeh, E.T.H., 2011. SUMOylation and de-SUMOylation in response to DNA damage. FEBS Letters 585, 2891–2896. doi:10.1016/j.febslet.2011.04.002
Drier, Y., Sheffer, M., Domany, E., 2013. Pathway-based personalized analysis of cancer. Proceedings of the National Academy of Sciences 110, 6388–6393. doi:10.1073/pnas.1219651110

Endesfelder, D., Burrell, R.A., Kanu, N., McGranahan, N., Howell, M., Parker, P.J., Downward, J., Swanton, C., Kschischo, M., 2014. Chromosomal instability selects gene copy-number variants encoding core regulators of proliferation in ER+ breast cancer. Cancer Research 74, 4853–4863. doi:10.1158/0008-5472.CAN-13-2664

Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N.J., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C., Martin, N.M.B., Jackson, S.P., Smith, G.C.M., Ashworth, A., 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434, 917–921. doi:10.1038/nature03445

Habermann, J.K., Doering, J., Hautaniemi, S., Roblick, U.J., Bündgen, N.K., Nicorici, D., Kronenwett, U., Rathnagiriswaran, S., Mettu, R.K.R., Ma, Y., Krüger, S., Bruch, H.-P., Auer, G., Guo, N.L., Ried, T., 2009. The gene expression signature of genomic instability in breast cancer is an independent predictor of clinical outcome. Int. J. Cancer 124, 1552–1564. doi:10.1002/ijc.24017

Hastings, P.J., Lupski, J.R., Rosenberg, S.M., Ira, G., 2009. Mechanisms of change in gene copy number. Nat Rev Genet 10, 551–564. doi:10.1038/nrg2593

Hu, Z., Fan, C., Oh, D.S., Marron, J.S., He, X., Qaqish, B.F., Livasy, C., Carey, L.A., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M.G., Sawyer, L.R., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Ruiz Orrico, A., Dreher, D., Palazzo, J.P., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J.F., Ellis, M.J., Olopade, O.I., Bernard, P.S., Perou, C.M., 2006. The molecular portraits of breast tumors are conserved across microarray platforms. BMC Genomics 7, 96. doi:10.1186/1471-2164-7-96

Huang, S., Yee, C., Ching, T., Yu, H., Garmire, L.X., 2014. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput Biol 10, e1003851. doi:10.1371/journal.pcbi.1003851

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28, 27–30.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al., 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. doi:10.1038/nature11412

Lee, A.J.X., Endesfelder, D., Rowan, A.J., Walther, A., Birkbak, N.J., Futreal, P.A., Downward, J., Szallasi, Z., Tomlinson, I.P.M., Howell, M., Kschischo, M., Swanton, C., 2011. Chromosomal instability confers intrinsic multidrug resistance. Cancer Research 71, 1858–1870. doi:10.1158/0008-5472.CAN-10-3604

Lehmann, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., Pietenpol, J.A., 2011. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. J. Clin. Invest. 121, 2750–2767. doi:10.1172/JCI45014

Liu, C., Srihari, S., Lê Cao, K.-A., Chenevix-Trench, G., Simpson, P.T., Ragan, M.A., Khanna, K.K., 2014. A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. Nucleic Acids Research 42, 6106–6127. doi:10.1093/nar/gku284

Livshits, A., Git, A., Fuks, G., Caldas, C., Domany, E., 2015. Pathway-based personalized analysis of breast cancer expression data. Molecular Oncology. doi:10.1016/j.molonc.2015.04.006

Martin, R.W., Orelli, B.J., Yamazoe, M., Minn, A.J., Takeda, S., Bishop, D.K., 2007. RAD51 up-regulation bypasses BRCA1 function and is a common feature of BRCA1-deficient breast tumors. Cancer Research 67, 9658–9665. doi:10.1158/0008-5472.CAN-07-0290

Negrini, S., Gorgoulis, V.G., Halazonetis, T.D., 2010. Genomic instability--an evolving hallmark of cancer. Nat Rev Mol Cell Bio 11, 220–228. doi:10.1038/nrm2858

Orthwein, A., Fradet-Turcotte, A., Noordermeer, S.M., Canny, M.D., Brun, C.M., Strecker, J., Escribano-Diaz, C., Durocher, D., 2014. Mitosis inhibits DNA double-strand break repair to guard against telomere fusions. Science 344, 189–193. doi:10.1126/science.1248024

Ottaviani, D., LeCain, M., Sheer, D., 2014. The role of microhomology in genomic structural variation. Trends Genet. 30, 85–94. doi:10.1016/j.tig.2014.01.001

Parker, J.S., Mullins, M., Cheang, M.C.U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenbush, J.F., Stijleman, I.J., Palazzo, J., Marron, J.S., Nobel, A.B., Mardis, E., Nielsen, T.O., Ellis, M.J., Perou, C.M., Bernard, P.S., 2009. Supervised risk predictor of breast cancer based on intrinsic subtypes. J. Clin. Oncol. 27, 1160–1167. doi:10.1200/JCO.2008.18.1370

Peng, G., Chun-Jen Lin, C., Mo, W., Dai, H., Park, Y.-Y., Kim, S.M., Peng, Y., Mo, Q., Siwko, S., Hu, R., Lee, J.-S., Hennessy, B., Hanash, S., Mills, G.B., Lin, S.-Y., 2014. Genome-wide transcriptome profiling of homologous recombination DNA repair. Nature communications 5, 3361. doi:10.1038/ncomms4361

Peng, Y., Dai, H., Wang, E., Lin, C.C.-J., Mo, W., Peng, G., Lin, S.-Y., 2015. TUSC4 functions as a tumor suppressor by regulating BRCA1 stability. Cancer Research 75, 378–386. doi:10.1158/0008-5472.CAN-14-2315

Perou, C.M., Sørlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lønning, P.E., Borresen-Dale, A.L., Brown, P.O., Botstein, D., 2000. Molecular portraits of human breast tumours. Nature 406, 747–752. doi:10.1038/35021093

Pitroda, S.P., Pashtan, I.M., Logan, H.L., Budke, B., Darga, T.E., Weichselbaum, R.R., Connell, P.P., 2014. DNA repair pathway gene expression score correlates with repair proficiency and tumor sensitivity to chemotherapy. Sci Transl Med 6, 229ra42. doi:10.1126/scitranslmed.3008291

Popova, T., Manié, E., Rieunier, G., Caux-Moncoutier, V., Tirapo, C., Dubois, T., Delattre, O., Sigal-Zafrani, B., Bollet, M., Longy, M., Houdayer, C., Sastre-Garau, X., Vincent-Salomon, A., Stoppa-Lyonnet, D., Stern, M.-H., 2012. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. Cancer Research 72, 5454–5462. doi:10.1158/0008-5472.CAN-12-1470

Price, B.D., D'Andrea, A.D., 2013. Chromatin remodeling at DNA double-strand breaks. Cell 152, 1344–1354. doi:10.1016/j.cell.2013.02.011

Przybytkowski, E., Lenkiewicz, E., Barrett, M.T., Klein, K., Nabavi, S., Greenwood, C.M.T., Basik, M., 2014. Chromosome-breakage genomic instability and chromothripsis in breast cancer. BMC Genomics 15, 579. doi:10.1186/1471-2164-15-579

Richardson, C., Stark, J.M., Ommundsen, M., Jasin, M., 2004. Rad51 overexpression promotes alternative double-strand break repair pathways and genome instability. Oncogene 23, 546–553. doi:10.1038/sj.onc.1207098

Roylance, R., Endesfelder, D., Gorman, P., Burrell, R.A., Sander, J., Tomlinson, I., Hanby, A.M., Speirs, V., Richardson, A.L., Birkbak, N.J., Eklund, A.C., Downward, J., Kschischo, M., Szallasi, Z., Swanton, C., 2011. Relationship of extreme chromosomal instability with long-term survival in a retrospective analysis of primary breast cancer. Cancer Epidemiol. Biomarkers Prev. 20, 2183–2194. doi:10.1158/1055-9965.EPI-11-0343

Sansregret, L., Nepveu, A., 2011. Gene signatures of genomic instability as prognostic tools for breast cancer. Future Oncol 7, 591–594. doi:10.2217/fon.11.34

Sarasin, A., Kauffmann, A., 2008. Overexpression of DNA repair genes is associated with metastasis: a new hypothesis. Mutat Res 659, 49–55. doi:10.1016/j.mrrev.2007.12.002

Sharma, V., Misteli, T., 2013. Non-coding RNAs in DNA damage and repair. FEBS Letters 587, 1832–1839. doi:10.1016/j.febslet.2013.05.006

Smid, M., Hoes, M., Sieuwerts, A.M., Sleijfer, S., Zhang, Y., Wang, Y., Foekens, J.A., Martens, J.W.M., 2011. Patterns and incidence of chromosomal instability and their prognostic relevance in breast cancer subtypes. Breast Cancer Res Treat 128, 23–30. doi:10.1007/s10549-010-1026-5

Swanton, C., Nicke, B., Schuett, M., Eklund, A.C., Ng, C., Li, Q., Hardcastle, T., Lee, A., Roy, R., East, P., Kschischo, M., Endesfelder, D., Wylie, P., Kim, S.N., Chen, J.-G., Howell, M., Ried, T., Habermann, J.K., Auer, G., Brenton, J.D., Szallasi, Z., Downward, J., 2009. Chromosomal instability determines taxane response. Proceedings of the National Academy of Sciences 106,

8671–8676. doi:10.1073/pnas.0811835106
Sørlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lønning, P.E., Borresen-Dale, A.L., 2001. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc. Natl. Acad. Sci. U.S.A. 98, 10869–10874. doi:10.1073/pnas.191367098
Takata, M., Sasaki, M.S., Tachiiri, S., Fukushima, T., Sonoda, E., Schild, D., Thompson, L.H., Takeda, S., 2001. Chromosome instability and defective recombinational repair in knockout mutants of the five Rad51 paralogs. Molecular and Cellular Biology 21, 2858–2866. doi:10.1128/MCB.21.8.2858-2866.2001
Thompson, S.L., Bakhoum, S.F., Compton, D.A., 2010. Mechanisms of chromosomal instability. Curr Biol 20, R285–95. doi:10.1016/j.cub.2010.01.034
Turner, N., Tutt, A., Ashworth, A., 2004. Hallmarks of "BRCAness" in sporadic cancers. Nat Rev Cancer 4, 814–819. doi:10.1038/nrc1457
Ulrich, H.D., 2012. Ubiquitin and SUMO in DNA repair at a glance. Journal of Cell Science 125, 249–254. doi:10.1242/jcs.091801
van Attikum, H., Gasser, S.M., 2009. Crosstalk between histone modifications during the DNA damage response. Trends in Cell Biology 19, 207–217. doi:10.1016/j.tcb.2009.03.001
Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., Stuart, J.M., 2010. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics 26, i237–45. doi:10.1093/bioinformatics/btq182
Venkitaraman, A.R., 2014a. Cancer suppression by the chromosome custodians, BRCA1 and BRCA2. Science 343, 1470–1475. doi:10.1126/science.1252230
Venkitaraman, A.R., 2014b. Tumour suppressor mechanisms in the control of chromosome stability: insights from BRCA2. Mol. Cells 37, 95–99. doi:10.14348/molcells.2014.2346
Villarreal, D.D., Lee, K., Deem, A., Shim, E.Y., Malkova, A., Lee, S.E., 2012. Microhomology directs diverse DNA break repair pathways and chromosomal translocations. PLoS Genet 8, e1003026. doi:10.1371/journal.pgen.1003026
Vincent-Salomon, A., Benhamo, V., Gravier, E., Rigaill, G., Gruel, N., Robin, S., de Rycke, Y., Mariani, O., Pierron, G., Gentien, D., Reyal, F., Cottu, P., Fourquet, A., Rouzier, R., Sastre-Garau, X., Delattre, O., 2013. Genomic instability: a stronger prognostic marker than proliferation for early stage luminal breast carcinomas. PLoS ONE 8, e76496. doi:10.1371/journal.pone.0076496
Vollan, H.K.M., Rueda, O.M., Chin, S.-F., Curtis, C., Turashvili, G., Shah, S., Lingjaerde, O.C., Yuan, Y., Ng, C.K., Dunning, M.J., Dicks, E., Provenzano, E., Sammut, S., McKinney, S., Ellis, I.O., Pinder, S., Purushotham, A., Murphy, L.C., Kristensen, V.N., METABRIC Group, Brenton, J.D., Pharoah, P.D.P., Børresen-Dale, A.-L., Aparicio, S., Caldas, C., 2015. A tumor DNA complex aberration index is an independent predictor of survival in breast and ovarian cancer. Molecular Oncology 9, 115–127. doi:10.1016/j.molonc.2014.07.019
Wang, H., Cai, H., Ao, L., Yan, H., Zhao, W., Qi, L., Gu, Y., Guo, Z., 2015a. Individualized identification of disease-associated pathways with disrupted coordination of gene expression. Briefings in Bioinformatics. doi:10.1093/bib/bbv030
Wang, H., Sun, Q., Zhao, W., Qi, L., Gu, Y., Li, P., Zhang, M., Li, Y., Liu, S.-L., Guo, Z., 2015b. Individual-level analysis of differential expression of genes and pathways for personalized medicine. Bioinformatics 31, 62–68. doi:10.1093/bioinformatics/btu522
Watkins, J., Weekes, D., Shah, V., Gazinska, P., Joshi, S., Sidhu, B., Gillett, C., Pinder, S., Vanoli, F., Jasin, M., Mayrhofer, M., Isaksson, A., Cheang, M.C.U., Mirza, H., Frankum, J., Lord, C.J., Ashworth, A., Vinayak, S., Ford, J.M., Telli, M.L., Grigoriadis, A., Tutt, A.N.J., 2015. Genomic Complexity Profiling Reveals That HORMAD1 Overexpression Contributes to Homologous Recombination Deficiency in Triple-Negative Breast Cancers. Cancer Discov 5, 488–505. doi:10.1158/2159-8290.CD-14-1092
Wiechec, E., 2011. Implications of genomic instability in the diagnosis and treatment of breast

cancer. Expert Rev. Mol. Diagn. 11, 445–453. doi:10.1586/erm.11.21
Wilhelm, T., Magdalou, I., Barascu, A., Técher, H., Debatisse, M., Lopez, B.S., 2014. Spontaneous slow replication fork progression elicits mitosis alterations in homologous recombination-deficient mammalian cells. Proceedings of the National Academy of Sciences 111, 763–768. doi:10.1073/pnas.1311520111
Yanagawa, M., Ikemot, K., Kawauchi, S., Furuya, T., Yamamoto, S., Oka, M., Oga, A., Nagashima, Y., Sasaki, K., 2012. Luminal A and luminal B (HER2 negative) subtypes of breast cancer consist of a mixture of tumors with different genotype. BMC Res Notes 5, 376. doi:10.1186/1756-0500-5-376
Zeman, M.K., Cimprich, K.A., 2014. Causes and consequences of replication stress. Nat. Cell Biol. 16, 2–9. doi:10.1038/ncb2897
Zheng, L., Dai, H., Zhou, M., Li, X., Liu, C., Guo, Z., Wu, X., Wu, J., Wang, C., Zhong, J., Huang, Q., Garcia-Aguilar, J., Pfeifer, G.P., Shen, B., 2012. Polyploid cells rewire DNA damage response networks to overcome replication stress-induced barriers for tumour progression. Nature communications 3, 815. doi:10.1038/ncomms1825

Chapter Four: Integrating multi-omics data to dissect mechanisms of DNA repair dysregulation in breast cancer

In the preceding chapter, I showed that HR dysregulation is prevalent in breast cancer. Many previous studies on breast cancer also showed that some other DNA repair genes and/or pathways exhibit abnormal expression in this malignancy, and can be used as biomarkers to predict therapy response. All these observations raised an important question: what are the underlying mechanisms of DNA repair dysregulation in breast cancer?

In this chapter, I assessed the contributions of possible mechanisms, including DNA copy number alteration (CNA), DNA methylation at gene promoter regions and mRNA expression changes of transcription factors (TFs), to the differential expression of individual DNA repair genes in breast cancer. In particular, I developed a penalised linear regression-based statistical framework to identify TFs that are potentially associated with each differentially expressed DNA repair gene.

The results from this study indicate that CNA and expression changes of TFs are major contributors to DNA repair dysregulation in breast cancer, and that ten TFs, each of which has a number of targets in multiple DNA repair pathways, may exert a global impact on the repair dysregulation in this cancer type. This study thus provides new insights into the underlying mechanisms of DNA repair dysregulation in breast cancer. These insights improve our understanding of the molecular basis of the DNA repair biomarkers identified thus far, and have the potential to inform future biomarker discoveries.

These results are presented here as a manuscript in preparation. The supplementary data of this manuscript, which are all Excel files, are deposited at UQ eSpace.

Integrating Multi-omics Data to Dissect Mechanisms of DNA repair Dysregulation in Breast Cancer

Chao Liu[a], Florian Rohart[b], Peter T. Simpson[cd], Kum Kum Khanna[e], Mark A. Ragan[a*] and Kim-Anh Lê Cao[f*]

[a]Institute for Molecular Bioscience, The University of Queensland, St. Lucia, QLD 4067, Australia

[b]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, St. Lucia, QLD 4067, Australia

[c]The University of Queensland, UQ Centre for Clinical Research, Herston, QLD 4029, Australia

[d]School of Medicine, The University of Queensland, Herston, QLD 4006, Australia

[e]QIMR-Berghofer Medical Research Institute, Herston, Brisbane, QLD 4029, Australia

[f]University of Queensland Diamantina Institute, Translational Research Institute, Woolloongabba, QLD 4102, Australia

[*] Joint corresponding authors:

Kim-Anh Lê Cao, University of Queensland Diamantina Institute, Translational Research Institute, 37 Kent Street, Woolloongabba, QLD 4102, Australia. Tel: +61 7 3443 7069; Fax: +61 7 3443 6966; Email: k.lecao@uq.edu.au

Mark A. Ragan, Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Rd, St. Lucia, QLD 4067, Australia. Tel.: +61 7 3346 2616; Fax: +61 7 3346 2101; E-mail: m.ragan@uq.edu.au

Abstract

The cellular status of DNA repair mechanisms has important implications for both carcinogenesis and cancer treatment, as supported by the observations that mutations in key DNA repair genes predispose individuals to a wide variety of cancers, and DNA repair capacity is one of the decisive factors underlying the outcome of chemo- and radiotherapy. DNA repair genes and pathways that are transcriptionally dysregulated in cancer provide the first line of evidence for the altered DNA repair status in tumours, and hence have been explored intensively as a reservoir for biomarker discovery. The molecular mechanisms underlying DNA repair dysregulation, however, have not been systematically investigated in any cancer type. In this study, we dissected the roles of DNA copy number alteration (CNA), DNA methylation at gene promoter regions (DM) and the expression changes of transcriptional factors (TFs) in the differential expression of individual DNA repair genes in normal versus tumour breast samples. These gene-level results were summarised at pathway-level to assess whether different DNA repair pathways are affected in distinct manner. In particular, we developed a penalised linear regression-based statistical framework to identify TFs potentially involved in the dysregulation of individual DNA repair genes. Our results demonstrated that CNA and expression changes of TFs are major causes of DNA repair dysregulation in breast cancer, and that a subset of the identified TFs may have global impact on the dysregulation of multiple repair pathways. This study thus provides novel insights into DNA repair dysregulation in breast cancer. These insights improve our understanding of the molecular basis of the DNA repair biomarkers identified thus far, and have the potential to inform future biomarker discovery.

Keywords: *multi-omics data analysis*; *DNA repair*; *transcription factor; DNA copy number*; *DNA methylation*; *breast cancer*

Introduction

Cells have evolved complex mechanisms to repair DNA lesions that arise from various endogenous and exogenous factors, including ultraviolet radiation, chemical carcinogens and oxidative by-products from normal cellular respiration. Hundreds of DNA repair genes have been identified, which mainly participate in five distinct but functionally intermingled pathways: homologous recombination (HR), non-homologous end joining (NHEJ), nucleotide excision repair (NER), base excision repair (BER) and mismatch repair (MMR). The functionalities of these pathways and their constituent components have been elucidated in detail [1-5].

The cellular status of DNA repair genes or pathways has important implications for both carcinogenesis and cancer treatment. Mutations in DNA repair genes predispose to a wide variety of cancers, and altered DNA repair efficiency is one of the decisive factors underlying the response of DNA damage-based cancer therapies [1-5]. DNA repair genes and pathways that are transcriptionally dysregulated in tumours carry valuable information with regard to drug response, patient survival and tumour characteristics, and thus have been extensively studied [6-14]. For instance, Santarpia et al. [11] analysed the expression profiles of 145 DNA repair genes in untreated breast cancer patients versus breast cancer patients treated with chemotherapeutic agents. The authors found that the upregulation of nine genes (*BUB1*, *FANCI*, *MNAT1*, *PARP2*, *PCNA*, *POLQ*, *RPA3*, *TOP2A*, and *UBE2V2*) are associated with poor prognosis, and that of one gene (*ATM*) is associated with good prognosis [11]. At the pathway level, Kang et al. [10] devised a DNA repair pathway-focused score (DRPFS) by combining the expression levels of 23 genes involved in platinum-induced DNA damage repair; this DRPFS score outperforms other clinical factors in predicting treatment response of ovarian cancer patients [10]. More recently, our group [12] developed a homologous recombination (HR) score based on the expression of about 70 core HR genes. This score reflects HR repair efficiency and correlates with chromosomal instability as well as patient survival [12]. While the dysregulation of DNA repair genes and/or pathways has been documented in many studies, to our knowledge, the molecular mechanisms underlying these transcriptional abnormalities have not been systematically elucidated in any cancer type.

Cancer-related gene expression alterations may result from genetic and/or epigenetic changes in tumours, including DNA copy number alteration (CNA) and DNA methylation (DM) of CpG islands at gene promoter regions. In fact, aberrantly expressed genes with CNA or DM are more likely to be critical genes that drive tumour initiation and progression. For example, *MYC* was considered an oncogene candidate as its overexpression together with its copy-number gain were commonly observed in cancer [15], which led to subsequent experiments that further validated its oncogenic role [16]. Similarly, the tumour suppressor gene *PTEN* was often found to be

underexpressed in tumours that also exhibit *PTEN* deletion [17]. The recent availability of multi-omics data in several major cancer types has facilitated studies that aim to obtain a more-holistic understanding of the global impact of CNA or DM on the transcriptomic changes [18-21]. However, effects specific to DNA repair dysregulation have yet to be elucidated.

TFs are key cellular components that serve to activate or repress the transcription of their target genes, and as such are important mediators of many cellular pathways, including the DNA repair mechanisms. Expression changes of TF genes in cancer are often crucial events as they are frequently associated with tumour initiation and/or development. For example, a recent meta-analysis revealed that the transcriptional regulatory network in colorectal adenomas is characterised by more than 250 differentially expressed TF genes, a considerable fraction of which have established roles in colorectal tumourigenesis [22].

Identifying target genes for the individual TFs is challenging. Motif-based computational prediction of TF binding sites at gene promoter regions has long been used to infer TF-target relationship [23-26]; however, it is a well-known issue that such analyses tend to give false positive results mainly due to the short length of the motifs and lack of tissue specificity. In recent years, ChIP-Seq, which combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing, has been employed to produce genome-wide binding profiles for individual TFs in a cell line-specific manner. This technique can generate relatively accurate information about TF binding sites genome-wide, however, due to its high cost, as of now only a limited number of TFs have been profiled in certain cell lines [27]. Moreover, for TFs whose binding profiles have been measured by ChIP-Seq, defining their target genes still remains an open question [28].

Breast cancer is one of the most common malignancies worldwide. This malignancy has a particularly close relationship with DNA repair defects, with the two well-known breast cancer susceptibility genes, *BRCA1* and *BRCA2*, being essential components of the HR repair pathway [29,30]. Previous studies showed that DNA repair genes and/or pathways are frequently dysregulated in breast cancer [11-14]. In this study, we aimed to provide biological insights regarding the underlying mechanisms of DNA repair dysregulation in this cancer type, taking advantage of the multi-omics data recently generated by the Cancer Genome Atlas (TCGA) [20]. Towards this aim, we first identified DNA repair genes that are differentially expressed between normal and tumour breast samples. Next, we evaluated the in *cis* effects of CNA and DM on the expression alteration of the identified DNA repair genes. Finally, we developed a penalised linear regression-based statistical framework, which takes into account the effects of CNA and DM on gene expression, to identify TFs potentially associated with each differentially expressed DNA repair gene. Our results showed that CNA and the expression changes of the identified TFs can statistically explain most of the expression variance of the repair genes, indicating the potential

importance of these two factors in driving DNA repair dysregulation in this common type of malignancy.

Materials and Methods

1. Data collection

The preprocessed genomic data generated by TCGA [20], including gene expression data for 113 normal breast tissues, and gene expression, CNA and DM data for 720 breast tumour samples, were retrieved via the UCSC Cancer Genomics Browser (https://genome-cancer.ucsc.edu/). Gene expression data for both the normal tissue and tumour samples had been generated using the Illumina HiSeq 2000 RNA sequencing platform, and show the Expectation Maximisation (RSEM)-normalised and percentile-ranked gene-level transcription estimates. The CNA data had been produced using Affymetrix SNA 6.0 arrays, with germline copy-number variation filtered out. The CNA values we obtained are gene-level segmentation values where value 0 represents the diploid state of the chromosome. The DM profiles had been produced with the Illumina Infinium HumanMythylation450 platform. The preprocessed methylation values we obtained, known as beta values, are continuous variables between 0 and 1, representing the percentage of methylation at the gene promoter region (defined as regions from 1.5 kb upstream to 0.5 kb downstream of transcription start site).

The pre-processed ENCODE Chip-Seq data for seven TFs (E2F1, MYC, TCF7L2, CTCF, GATA3, ZNF217 and POLR2A) measured on the breast cancer cell line MCF-7 were downloaded from the UCSC genome browser (https://genome.ucsc.edu/encode/). Detailed information about how these data had been generated, including antibody and immunoprecipitation specificity, library complexity, sequencing depth, peak calling and quality assessment, were elaborated in a dedicated publication by ENCODE [31]. Briefly, in each ChIP-Seq experiment for mammalian genomes, ENCODE generated ≥ 10 million uniquely mapping reads (25-36 bp in length), providing a reliable and valuable data source for TF studies [31].

2. Differential expression analysis

DNA repair genes differentially expressed in tumour versus normal breast tissues were identified using Limma [32], with the criterion that false discovery rate (FDR) < 0.05 after Benjamini and Hochberg's multiple-test adjustment [33].

3. Identification of TFs potentially involved in DNA repair dysregulation

To systematically search for TFs potentially involved in DNA-repair dysregulation in breast cancer, we performed TF selection from a list of 1391 manually curated human TFs, which was estimated to cover 85% to 94% of all human TFs [34]. Specifically, for each differentially expressed DNA repair gene identified in this study, we built a linear regression model connecting CNA, DM and the transcriptional changes of the 1391 TFs to explain the observed expression variance of the repair gene. For each repair gene *g*, we formulate the model as:

$$\mathcal{Y}_g = \beta_{0,g} + \beta_{cn,g}\mathcal{X}_{cn,g} + \beta_{Me,g}\mathcal{X}_{Me,g} + \sum_{k=1,\ldots,1391} \beta_{TF,k,g}\mathcal{X}_{TF,k} + \varepsilon_g$$

where $\mathcal{Y}_g$, $\mathcal{X}_{cn,g}$ and $\mathcal{X}_{Me,g}$ represent the abundances of mRNA, CNA, and DM of DNA repair gene *g*, respectively, while $\mathcal{X}_{TF,k}$ denotes the mRNA level of TF *k*. The regression coefficients $\beta_{cn,g}$ and $\beta_{Me,g}$ estimate the in *cis* contributions of CNA and DM to the expression changes of the repair gene *g*, while $\beta_{TF,k,g}$ evaluates the influence of the transcriptional changes of TF *k* on the expression changes of repair gene *g*. The intercept is represented by $\beta_{0,g}$ and error term by ε_g.

We then applied LASSO penalisation constraints [35] through the R package *glmnet* [36] to select a subset of the 1391 TFs whose transcriptional changes are significantly associated with the dysregulation of a given DNA repair gene. To ensure that the effects of CNA and DM on gene expression are always taken into consideration, we imposed an additional constraint, also through the *glmnet* package, that the regression coefficients for CNA and DM are never set to zero by LASSO during this feature selection process.

In practice, a major drawback of LASSO is that its result can be heavily affected by an initiating parameter termed regularisation coefficient (λ), whose value needs to be specified for each analysis. For a given analysis, this value is typically obtained using cross-validation; however, due to the randomness associated with the cross-validation process, the estimated optimal λ value for the same analysis can differ across different cross-validation runs, resulting in unstable feature selection results. To overcome this, we developed a secondary feature-selection procedure with the assumption that TFs consistently selected with different λ values are likely to be truly associated with a given repair gene (Figure 1).

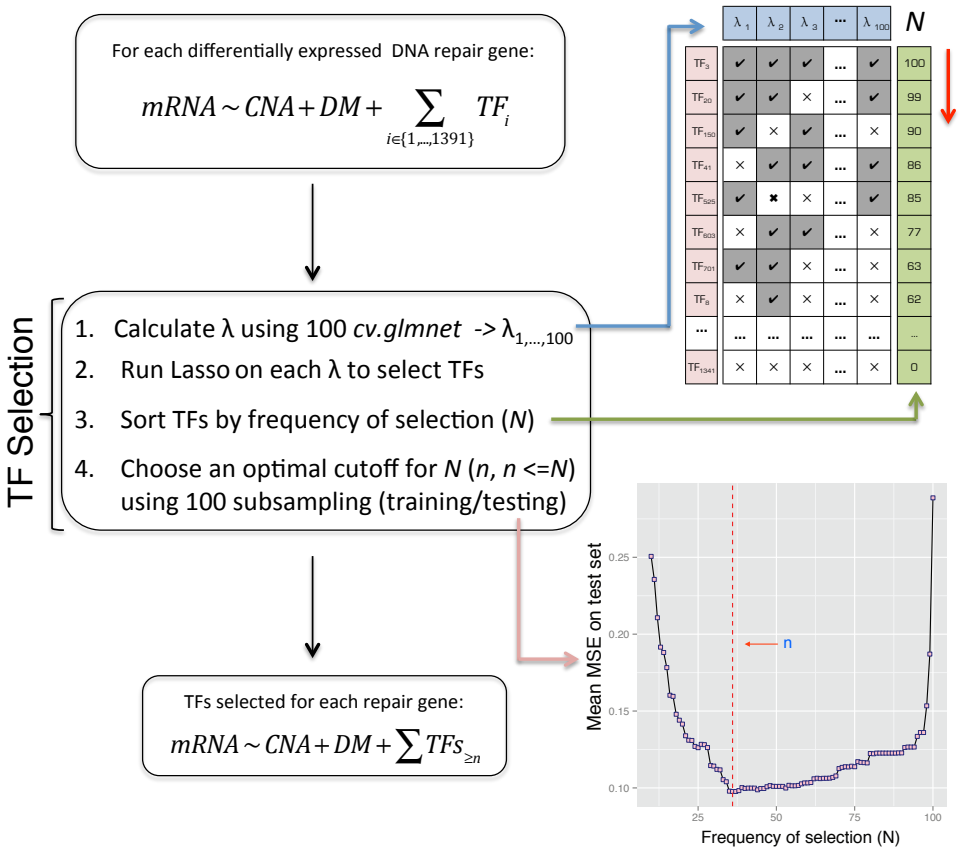


Figure 1 A LASSO-based statistical framework to identify TFs potentially involved in DNA repair dysregulation.

Specifically, for each differentially expressed DNA repair gene, we generated 100 different λ values by running the *cv.glmnet* function 100 times, and then performed LASSO on each λ. This led to 100 different, but overlapping, sets of TFs selected for the same repair gene. Next, we calculated for each TF the frequency of being selected across the 100 LASSO runs. This frequency, denoted as N ($1 \leq N \leq 100$), is important as it indicates the selection stability for each TF. To determine an optimal cutoff for N, we further built different regression models, whose response and explanatory variables are similar to the one described above except that, instead of including all the 1391 TFs, only the TFs that had been selected more than N times were included. As different values of N correspond to different sets of TFs, and in turn to different regression models, we reasoned that the optimal cutoff of N could be obtained by comparing the performance of these models. For this purpose, in the following step we randomly divided the samples into a training set (2/3 of all samples) and a testing set (the remaining 1/3 of all samples) for model training and testing, respectively. We repeated this subsampling process 100 times, and each time the performance of each model on the testing set was recorded as the mean squared error (MSE). The value of N that gave the minimal averaged MSE across the 100 subsampling was considered the optimal cutoff, and the TFs whose selection frequency was above this cutoff were considered to be associated with a given DNA repair gene (Figure 1).

4. Four alternative linear regression models to dissect the contributions of possible mechanisms to DNA repair dysregulation

To dissect the contributions of CNA, DM and TF-gene expression change to DNA repair gene dysregulation, for each differentially expressed DNA repair genes we further constructed four alterative linear regression models. Each model uses the mRNA abundance of the same DNA repair gene as response variable, but comprises different explanatory variables as follows:

(1) CNA + DM + TFs

(2) CNA + DM

(3) DM only

(4) CNA only

We compared the performance of these four alternative models via a subsampling-based process. Specifically, we randomly selected two-thirds of the tumour samples to train each of the four models, and the remaining one-third of the samples was used for testing the model performance. This process was repeated 100 times for each differentially expressed DNA repair gene, and the average performance of each model on the testing set was recorded.

Results

1. Identification of DNA repair genes that are differentially expressed between tumour and normal breast tissues

We manually curated a list of 195 DNA repair genes (Supplementary Table 1) by systematic literature search and consultation with a domain expert. This list includes genes from the five major DNA repair pathways and the Fanconi anaemia (FA) pathway, which is responsible for the repair of DNA inter-strand crosslinks and is closely associated with breast cancer susceptibility [37]. Of these 195 genes, 169 have CNA, DM and expression data in TCGA, of which 149 (88%) are differentially expressed between normal and tumour breast samples (Table 1; see Supplementary Table 2 for detail). This high percentage of differential expression is consistent with the existing knowledge that DNA repair genes are frequently dysregulated in breast cancer. Of the 149 differentially expressed repair genes, 106 (71%) exhibit significantly increased expression, and 43 (29%) show reduced expression. Similar observations were obtained when the number of up- and down-regulated genes within each individual repair pathway was examined separately (Table 1), indicating that DNA repair genes are more likely to be up-regulated than down-regulated in breast cancer.

Table 1 Number of differentially expressed (DE) genes in each DNA repair pathway

Pathway	Curated*	Present In TCGA	DE	Overexpressed	Underexpressed
HR	82	60	60	43	17
NER	66	48	48	36	12
BER	31	27	27	24	3
FA	31	23	23	19	4
NHEJ	25	22	22	13	9
MMR	24	20	20	15	5
Total†	195	169	149	106	43

* This column represents the number of manually curated genes in each repair pathway.
† Genes that appear in two or more pathways were counted only once.

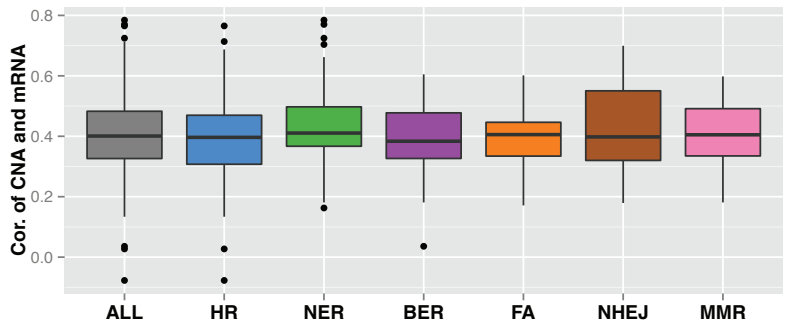
2. Estimate the effects of genetic and epigenetic changes on the DNA repair dysregulation in breast cancer

2.1. Contribution of CNA to the DNA repair dysregulation

To evaluate the effects of CNA and DM on DNA repair dysregulation, for each of the differentially expressed genes identified above, we measured the respective correlations of mRNA with CNA and DM using Spearman correlation coefficients. Figure 2 summarises the results for the correlations between CNA and mRNA, and Figure 3 for the correlations between DM and mRNA.

As shown in Figure 2A, the correlations between CNA and mRNA are in general modest, with a median correlation coefficient of about 0.4 (see Supplementary Table 3 for detail). As a background, the median correlation for the other 16,946 genes with CNA and mRNA data but not involved in DNA repair is 0.23. Out of the 149 differentially expressed DNA repair genes, 148 show positive correlations between CNA and mRNA, of which 146 have significant correlations (FDR < 0.05; Supplementary Table 3). These positive correlations are consistent with the role of CNA in modulating gene expression, and indicate that CNA plays an important role in driving DNA repair dysregulation in breast cancer. Similar patterns of modest positive correlations were observed when either all differentially expressed genes were considered, or when only genes within each repair pathway were included (Figure 2A), indicating that CNA affects different repair pathways in a similar way.

A



B

Top10 DNA repair genes with the highest CNA-mRNA correlation

Gene	Cor.	p-value	Pathway	Expression
POLR2K	0.78	6.59E-149	NER	Up
POLR2C	0.77	1.86E-140	NER	Down
CSNK2A2	0.77	9.37E-138	HR	Down
ERCC5	0.72	1.21E-116	NER	Down
RNF40	0.71	1.41E-111	HR	Up
CUL4A	0.70	2.64E-107	NER	Up
XRCC6	0.70	1.51E-105	NHEJ	Up
RAD54B	0.69	1.65E-100	HR	Up
TCEA1	0.66	9.51E-91	NER	Up
PARP1	0.60	6.68E-72	HR, NHEJ, BER	Up

Figure 2 The effect of CNA on DNA repair gene expression. (A) Distributions of the in *cis* correlations between CNA and mRNA expression, summarised for all differentially expressed DNA repair genes, or only genes from each individual repair pathway. (B) The top ten DNA repair genes, sorted by their in *cis* correlation between CNA and mRNA expression.

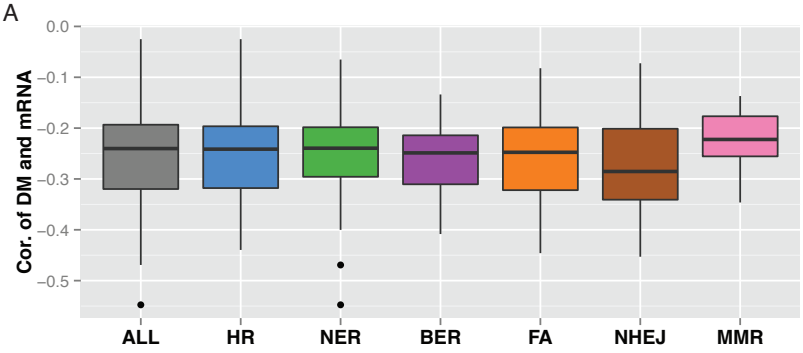
Figure 2B displays the top ten DNA repair genes whose differential expression is most likely due to their altered copy number status (i.e., these ten genes have the highest correlations between CNA and mRNA). For example, the up-regulation of *POLR2K* can be largely ascribed to its copy number gain while the down-regulation of *POLR2C* is mainly due to its copy number loss. We consider that these relative high correlations between inherent genetic changes and differential expression may have important implications for breast cancer therapy. For instance, recently studies showed that the *CUL4A* is a promising biomarker for a variety of cancers, including breast caner, whose overexpression is associated with elevated drug sensitivity [38,39]; and here we revealed that *CUL4A* overexpression in breast cancer is mainly induced by its copy number gain (Figure 2B). As another example, the protein encoded by *PARP1* is a newly proposed drug target in breast cancer [40], and here we showed that there is a relative high correlation between its mRNA overexpression and DNA amplification (Figure 2B).

2.2. Contribution of DM to the DNA repair dysregulation

Compared to the correlations between CNA and mRNA, the correlations between DM and mRNA are in general weak, with a median value of about -0.25 (Figure 3A and Supplementary Table 3; the median correlation for the other 13,382 genes with DM and mRNA data but not involved in DNA repair is -0.27). This is the case both for all differentially expressed genes, and for only those genes within each individual repair pathway (Figure 3A). Nonetheless, all 149

differentially expressed DNA repair genes show negative correlation between DM and expression, and 143 of them are significant (FDR < 0.05; Supplementary Table 3). This is in accordance with the role of methylation in suppressing gene expression. We found the DM-mRNA correlations are not significantly different between the up-regulated genes and down-regulated genes (p-value = 0.5, Wilcoxon rank-sum test), suggesting that DM is not a major factor for the downregulation of DNA repair genes in breast cancer. This observation is in line with a recent meta-analysis showing cancer-specific methylation patterns usually have marginal effects on mRNA expression [41].

A



B

Top10 DNA repair genes with the highest DM-mRNA correlations

Gene	Cor.	p-value	Pathway	Expression
TCEA3	-0.55	2.16E-55	NER	Up
TCEA1	-0.47	8.09E-39	NER	Up
PARP3	-0.45	5.04E-36	NHEJ	Down
FANCA	-0.45	6.76E-35	FA	Up
RAD54B	-0.44	6.62E-34	HR	Up
PSIP1	-0.43	1.98E-32	HR	Down
WRN	-0.42	2.65E-31	NHEJ	Down
MUTYH	-0.41	4.85E-29	BER	Up
POLB	-0.41	6.00E-29	BER	Up
ERCC5	-0.40	6.83E-28	NER	Down

Figure 3 The effect of DM on DNA repair gene expression. (A) Distributions of the in *cis* correlations between DM and mRNA expression, summarised for all differentially expressed DNA repair genes, or only genes from each individual repair pathway. (B) The top ten DNA repair genes, sorted by their in *cis* correlation between DM and mRNA expression.

A few DNA repair genes have modest correlations between DM and mRNA (Figure 3B). These genes are not enriched with downregulated repair genes (p-value =1, Fisher's exact test), indicating again that DM is not the major cause for reduced expression of DNA repair genes in breast cancer. However, DM may have important effect on the underexpression of some DNA repair genes. For example, of the downregulated repair genes listed in Figure 3B, the transcriptional silencing of *WRN* by promoter hypermethylation has been frequently observed in a number of cancers, including breast cancer [42]. This epigenetic inactivation of *WRN* can lead to increased chromosomal instability and hypersensitivity to DNA-damaging drugs, and thus has important implications for cancer therapy [42,43].

3. Estimate the effect of TF transcriptional changes on the DNA repair dysregulation in breast cancer

3.1. TFs identified by our LASSO-based statistical framework and their contribution to the DNA repair dysregulation

One of the aims of this work was to systematically search for TFs that are likely to be involved in DNA repair dysregulation in breast cancer. To this end, we first downloaded a list of 1391 manually curated TFs, which was estimated to cover 85% to 94% of all the human TFs [34]. Next, for each of the 149 differentially expressed DNA repair genes identified above, we built a linear regression model connecting CNA, DM and the transcriptional changes of the 1391 TFs to explain the observed repair gene dysregulation. Since the vast majority of the 1391 TFs are not associated with the dysregulation of a particular repair gene, we further developed a LASSO-based statistical framework to identify relevant TFs for each repair gene (see Materials and Methods for detail).

In brief, the LASSO constraint [35] enforces scarcity in a linear regression model (i.e., enforcing most of the small regression coefficients to be zero) and thus reduces the number of explanatory variables included in the model. To account for the effects of CNA and DM on gene expression, we imposed an additional constraint that the regression coefficients of CNA and DM will never be set to zero by LASSO. In other words, after taking into account the confounding effects from CNA and DM, we identified TFs whose transcriptional changes are associated with the aberrant expression of each DNA repair gene. Through this approach, we identified 6 to 132 relevant TFs (with a median value of 39) for each differentially expressed DNA repair gene (Supplementary Table 4). Many of these TFs have established roles in DNA repair, and some of them are discussed in the next section.

To dissect the contributions of CNA, DM and TF-gene expression change to DNA repair gene dysregulation, next we constructed four alternative linear regression models for every differentially expressed DNA repair gene. Each model uses the mRNA abundance of the same DNA repair gene as the response variable, but comprises different explanatory variables as shown in Table 2. We compared the performance of the four models for the same repair gene via a subsampling-based process (see Materials and Methods for detail), and summarised the results across the 149 differentially expressed DNA repair genes in terms of two measurements: Spearman correlation coefficient between predicted and observed mRNA abundance, and variance in the mRNA abundance explained by the model (coefficient determination, R^2).

As we can see from Table 2, the model using CNA alone as explanatory variable performs better than the model using DM alone (average Spearman correlation coefficient 0.41 vs 0.25, and

R^2, 22% vs 0%), which is consistent with the result from Section 1 showing that CNA in general has a higher correlation with mRNA than DM. Furthermore, results in Table 2 also reveal that, compared with using CNA alone, combining CNA and DM does not greatly improve the model performance (average Spearman correlation coefficient 0.44 vs 0.41, and R^2, 24% vs 22%). By contrast, when the expression values of the selected TFs are added, the model performance becomes substantially improved (average Spearman correlation coefficient 0.85 vs 0.44, and R^2, 73% vs 24%). These results demonstrate that using the expression values of the identified TFs can substantially improve the model performance, which underscores the importance of these TFs in driving DNA repair dysregulation.

Table 2 Performance comparison of the four linear regression models. Each model uses the mRNA abundance of the same DNA repair gene as the response variable, but comprises different explanatory variables as listed below. A negative R^2 means that the linear model poorly fits the data.

	Spearman Correlation Coefficient (%)				Coefficient of Determination (R^2) (%)			
	Min	**Median**	**Mean**	**Max**	**Min**	**Median**	**Mean**	**Max**
DM	-13	24	25	55	-237	3	0	24
CNA	-7	40	41	78	-14	20	22	61
CNA + DM	-6	43	44	78	-13	21	24	61
TFs	37	77	75	95	12	56	55	87
CNA + DM + TFs	64	86	85	97	34	74	73	91

3.2. Selected TFs that may be major drivers of DNA repair dysregulation

Among the TFs identified by the LASSO-based statistical framework, some are predicted to target multiple genes within the same repair pathway, and therefore might be particularly important for the dysregulation of that pathway. Moreover, they may also target genes that function in different repair pathways, and hence may be able to exert a global influence on the dysregulation of the DNA repair machinery. With these thoughts in mind, we sorted the TFs identified in this study according to the number of genes that they target, and showed the top ten TFs and their pathway-specific targets in Figure 4. We consider these TFs as potential major drivers of DNA repair dysregulation in breast cancer.

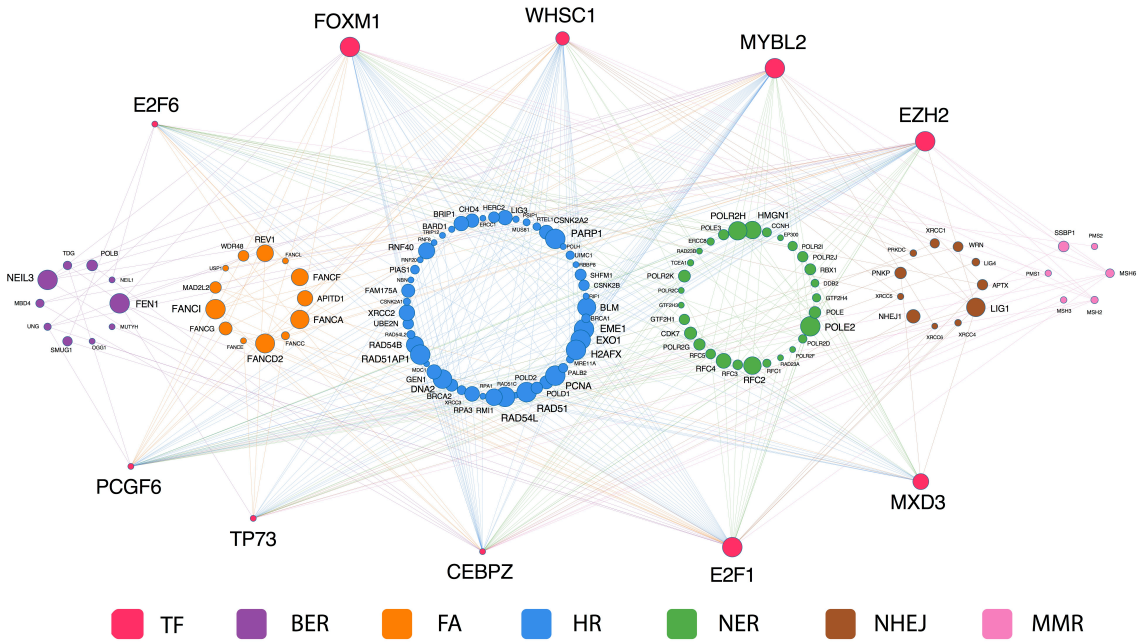


Figure 4 Ten TFs as potential major drivers of DNA repair dysregulation in breast cancer. TFs selected by the LASSO-based statistical framework were sorted by the number of target genes, and only the top ten TFs and their target genes are shown. The target genes are grouped according to pathway participation. Node size indicates level of differential expression. The figure was made using the Cytoscape software (http://www.cytoscape.org/) with manual layout.

Of these selected TFs, some have well-established roles in modulating DNA repair. The most prominent example is FOXM1, a master regulator of DNA damage response and a determinant of resistance to DNA-damaging agents [44]. Overexpression of the *FOXM1* gene is observed in many cancers [45], including breast cancer [46], and is thought to cause genomic instability [47] and poor prognosis [48,49]. Another noted DNA repair regulator is E2F1, which coordinates the function of several vital cellular processes, including DNA repair, cell cycle checkpoint and apoptosis [50-52]. A recent study showed that, following treatment with histone deacetylase inhibitors (HDACs), a promising class of drug in prostate cancer, decreased recruitment of E2F1 results in downregulation of a few key DNA repair genes, leading to reduced DNA repair capacity and enhanced sensitivity to genotoxic agents [53]. Interestingly, most of these key repair genes, including *BRCA1*, *RAD51*, *RAD54L* and *BLM*, were also identified in this study as E2F1 targets in breast cancer.

Apart from TFs with well-established role in DNA repair, the top ten TFs shown in Figure 4 also include TFs whose role in DNA repair is less-well studied. For example, the protein p73 (also known as TP73), which belongs to the same family as the well-known tumour suppressor p53, was recently discovered to regulate DNA repair gene expression [54]. As another example, MXD3, whose role in human DNA repair has not begun to be explored, was recently proposed to be involved in DNA repair in mouse [55]. We hence propose that these less well-known TFs identified in this study may serve as good candidates for identifying novel regulators of DNA repair and/or innovative drug targets for DNA repair-related breast cancer therapy.

3.3. TFs with ChIP-Seq profiles in ENCODE

In the LASSO-based statistical model, TFs were selected based on an association of the expression of the TF genes with the expression of a given DNA repair gene. One issue associated with this process is that some of the TFs selected for a given DNA repair gene may not directly regulate the repair gene, i.e. these TFs may function as upstream regulators of DNA repair, which do not directly bind and target a particular repair gene. We therefore sought to estimate the percentages of the identified TFs that have direct targets in DNA repair.

A major difficulty is that the genome-wide binding sites of most human TFs are currently unknown. For example, the Encyclopedia of DNA Elements (ENCODE) project, which aims to build a comprehensive list of functional elements in the human genome [56], describes only 161 TFs (~10% of all human TFs) that have ChIP-Seq data. These 161 TFs were profiled in 91 cell types, with each cell type having a few to dozens of TFs analysed (https://genome.ucsc.edu/encode/). In addition, for TFs whose binding sites have been measured by ChIP-Seq, how to define their direct target genes is still an open question [28].

Here we searched the ENCODE database for TFs which were identified in this study and also have binding profiles measured by ChIP-Seq. As all ENCODE ChIP-Seq data were measured in cell lines, here we used the breast cancer cell line MCF-7 as a surrogate for the TCGA breast cancer samples analysed in this study. MCF-7 has been widely used in breast cancer research, and it has more TFs measured by ChIP-Seq than do other breast cancer cell lines in ENCODE. Of the seven TFs measured in MCF-7, we found six (E2F1, MYC, TCF7L2, CTCF, GATA3, ZNF217) were identified in this study as potential DNA repair regulators. For each of these six TFs, we further examined how many of the predicted targets have support from the ChIP-Seq data. Specifically, we calculated the physical distances between TF binding sites and the transcription start sites (TSSs) of target genes located on the same chromosome; we consider a direct TF-target relationship to exist if such a distance is $\leq$ 100 kb (the criterion was chosen according to [57]). As shown in Table 3, although the small sample size used in this analysis may lead to a biased result, we found in total 81% of the predicted DNA repair targets are supported by the ChIP-Seq data; and in particular, of the 46 predicted E2F1 targets, 41 (89%) have support from this ChIP-Seq analysis. This result suggests that most of the TFs identified in this study are likely to directly regulate their DNA repair targets.

Table 3 TFs with predicted DNA-repair targets and ChIP-Seq profiles from ENCODE

TF	No. of predicted targets	No. of predicted targets supported by ChIP-Seq data	Percentage
E2F1	46	41	89%
MYC	12	10	83%
TCF7L2	8	5	63%
CTCF	6	6	100%
GATA3	6	2	33%
ZNF217	2	1	50%
Total	80	65	81%

Discussion

Prognostic and predictive biomarkers selected from high-throughput genomic data, which allow stratification of patients for tailored therapy, are of critical importance in cancer management [58]. Cancer-related dysregulation of DNA repair genes or pathways reflects altered DNA repair efficiency in tumours, and hence has been investigated intensively for biomarker discovery; to our knowledge, however, the genetic underpinnings of DNA repair dysregulation have not been systematically elucidated for any cancer type. In this study, we dissected gene-specific contributions of CNA, DM and expression changes of TFs to the differential expression of DNA repair genes between tumour and normal breast samples. We showed that CNA and expression changes of TFs are major causes of DNA repair dysregulation in breast cancer, and identified ten TFs that may potentially exert global impact on the dysregulation of multiple DNA repair pathways in this cancer type. Our work thus provides novel biological insights into DNA repair dysregulation in breast cancer. These insights improve our understanding of the molecular basis of the DNA repair biomarkers identified thus far, and have potential to inform future biomarker discovery.

Access to multi-omics data for major cancer types has been greatly facilitated by large-scale projects such as TCGA in recent years. Accordingly, many methods for integrative multi-omics data analysis have emerged, aiming to help us understand the interplay between different molecular levels, and/or provide improved power to identify important genomic factors [59,60]. Compared to other integrative methods, linear regression models have two distinct advantages for studying the altered transcriptional programs in cancer: 1) they regard the expression of a gene as a function of CNA, DM and TF activities etc., and thus provide a priming biological knowledge-based causal framework for data integration and gene expression modelling; and 2) unlike most integrative methods, which may encounter the “curse of dimensionality” when adding more data types into the analysis, linear regression models are quite flexible in this regard because even with a large number

of potential explanatory variables, a parsimonious model can still be obtained through penalisation (e.g, by LASSO).

Studies that utilise linear regression models for multi-omics data analysis have been reported recently [61-64]. For instance, Li and colleagues [61] fitted a linear model on the expression of each gene in acute myeloid leukaemia (AML) using gene-specific CNA, DM, TF binding signals and the counts of miRNA binding sites at the 3'-UTR as explanatory variables. In another study, Setty et al. [62] modelled gene expression change in glioblastoma as a linear function of CNA, DM, the number of TF binding sites at the promoter region, and the number of miRNA binding sites at the 3'-UTR. These studies successfully identified a dozen TFs and miRNAs as key drivers of global transcriptional changes in AML and glioblastoma, respectively [61,62].

The above-mentioned regression-based integrative analyses also have certain limitations. For example, while LASSO-enhanced linear regression models can achieve better prediction accuracy and interpretability by effectively reducing the number of explanatory variables needed, the variable selection results may be significantly influenced by the choice of the initiating factor λ. As a common practice in the field, such as in the aforementioned study in AML [61], the value of λ is determined by running a cross-validation function only once, which may lead to an unstable result due to the random nature of the cross-validation process. In this study, we addressed this issue by developing a secondary feature selection procedure that ensures the robustness of the TFs identified in this study (See Material and Methods for detail).

Another problem is associated with insufficiency of explanatory variables. For example, the number of TFs covered by either of the two above-mentioned studies was quite limited. Specifically, Li and colleagues [61] conducted TF identification from 97 TFs whose binding profiles were measured in K562, a cell line that by far has the highest number of TFs measured by ENCODE ChIP-Seq experiments; the TF binding information utilised by Setty et al. [62], retrieved from the TRANSFAC database [65], was available for only 152 TFs. Although TF-binding information from ChIP-Seq experiments or the TRANSFAC database can be more accurate, the vast majority of human TFs were nonetheless omitted from these studies. By contrast, in the current study we performed TF selection from a list of 1391 TFs, covering 85% to 94% of all human TFs. This high coverage enabled us to identify TFs potentially involved in DNA repair.

There are some remaining issues in this study that are mainly associated with lack of datasets. Firstly, our current model does not as yet consider the impact of miRNAs on gene expression, i.e. key variables may be absent. In fact, we failed to establish an association between dysregulation of DNA repair gene and expression changes of miRNAs (data not shown). The reason might be that in comparison with TFs, miRNAs usually have much smaller effects on target gene expression [66], and so given the large number of TFs in the model, miRNA-mediated

downregulation was not recognised by our LASSO-based approach. Future studies employing other features of miRNA, and/or other genomic datasets are needed to further refine the current model. Secondly, due to the lack of other large breast cohorts measured at multiple molecular levels, we had to perform model training and testing on the same TCGA breast cancer dataset. The availability of an independent dataset could provide more-accurate assessment for model performance. This limitation is common to many studies [61,62,64]. Thirdly, the TFs selected in this study were mainly based on statistical analysis and thus may contain false positives. Although the results in Section 3.3 indicate that our result may enjoy high accuracy, we hope that in the future our results can be further evaluated with against more experimentally based TF-binding profiles.

In summary, we developed a penalised regression-based statistical framework that can integrate CNA, DM and expression changes of TFs to explain DNA repair dysregulation in breast cancer. Our results demonstrated that CNA and TF expression changes are major factors affecting the dysregulation of individual DNA repair genes, and pointed to ten TFs that might be potential master drivers of DNA repair pathway dysregulation in this malignancy. This work thus facilitates our mechanistic understanding of how the exquisite control of DNA repair regulation is pathologically altered in breast cancer, and may provide important implications for future biomarker discovery. With the accumulation of ever-increasing amount of genomic data and developments in integrative analysis methods, a complete understanding of the transcriptional dysregulation in cancer will no longer beyond reach.

Acknowledgments

We thank TCGA for providing the genomic data. This study was funded by the Australian National Health and Medical Research Council (NHMRC) Project Grant (ID: 1028742) to PTS and MAR. KALC was supported in part by the Australian Cancer Research Foundation (ACRF) for the Diamantina Individualised Oncology Care Centre at The University of Queensland Diamantina Institute and NHMRC Career Development fellowship (1087415). KKK is an NHMRC Senior Principal Search Fellow (613638) supported by NHMRC Project Grant 1017028).

References

1. Curtin NJ. DNA repair dysregulation from cancer driver to therapeutic target. Nature Reviews Cancer. 2012;12: 801–817. doi:10.1038/nrc3399

2. Bouwman P, Jonkers J. The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. Nature Reviews Cancer. 2012;12: 587–598.

doi:10.1038/nrc3342

3. Lord CJ, Ashworth A. The DNA damage response and cancer therapy. Nature. 2012;481: 287–294. doi:10.1038/nature10760

4. Liu C, Srihari S, Lê Cao K-A, Chenevix-Trench G, Simpson PT, Ragan MA, et al. A fine-scale dissection of the DNA double-strand break repair machinery and its implications for breast cancer therapy. Nucleic Acids Research. 2014;42: 6106–6127. doi:10.1093/nar/gku284

5. Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities within the DNA damage response. Nature Reviews Cancer. 2015;15: 166–180. doi:10.1038/nrc3891

6. Soria-Bretones I, Sáez C, Ruíz-Borrego M, Japón MA, Huertas P. Prognostic value of CtIP/RBBP8 expression in breast cancer. Cancer Medicine. 2013;2: 774–783. doi:10.1002/cam4.141

7. Rezano A, Kuwahara K, Yamamoto-Ibusuki M, Kitabatake M, Moolthiya P, Phimsen S, et al. Breast cancers with high DSS1 expression that potentially maintains BRCA2 stability have poor prognosis in the relapse-free survival. BMC Cancer. 2013;13: 562. doi:10.1186/1471-2407-13-562

8. Kassambara A, Gourzones-Dmitriev C, Sahota S, Rème T, Moreaux J, Goldschmidt H, et al. A DNA repair pathway score predicts survival in human multiple myeloma: the potential for therapeutic strategy. Oncotarget. 2014;5: 2487–2498.

9. Pitroda SP, Pashtan IM, Logan HL, Budke B, Darga TE, Weichselbaum RR, et al. DNA repair pathway gene expression score correlates with repair proficiency and tumor sensitivity to chemotherapy. Science Translational Medicine. 2014;6: 229ra42. doi:10.1126/scitranslmed.3008291

10. Kang J, D'Andrea AD, Kozono D. A DNA Repair Pathway-Focused Score for Prediction of Outcomes in Ovarian Cancer Treated With Platinum-Based Chemotherapy. Journal of the National Cancer Institute. 2012;104: 670–681. doi:10.1093/jnci/djs177

11. Santarpia L, Iwamoto T, Di Leo A, Hayashi N, Bottai G, Stampfer M, et al. DNA repair gene patterns as prognostic and predictive factors in molecular breast cancer subtypes. Oncologist. 2013;18: 1063–1073. doi:10.1634/theoncologist.2013-0163

12. Liu C, Srihari S, Lal S, Gautier B, Simpson PT, Khanna KK, et al. Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. Molecular Oncology. 2015. doi:10.1016/j.molonc.2015.09.007

13. Aleskandarany M, Caracappa D, Nolan CC, Macmillan RD, Ellis IO, Rakha EA, et al. DNA damage response markers are differentially expressed in BRCA-mutated breast cancers. Breast Cancer Research and Treatment. 2015;150: 81–90. doi:10.1007/s10549-015-3306-6

14. Matta J, Morales L, Dutil J, Bayona M, Alvarez C, Suarez E. Differential expression of DNA repair genes in Hispanic women with breast cancer. Molecular Cancer Biology. 2013;1: 54. doi:10.9777/mcb.2013.10006

15. Marcu KB, Bossone SA, Patel AJ. myc function and regulation. Annual Review of

Biochemistry. 1992;61: 809–860. doi:10.1146/annurev.bi.61.070192.004113

16. Patel JH, Loboda AP, Showe MK, Showe LC, McMahon SB. Analysis of genomic targets reveals complex functions of MYC. Nature Reviews Cancer. 2004;4: 562–568. doi:10.1038/nrc1393

17. Song MS, Salmena L, Pandolfi PP. The functions and regulation of the PTEN tumour suppressor. Nature Reviews Molecular Cell Biology. 2012;13: 283–296. doi:10.1038/nrm3330

18. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486: 346-352. doi:10.1038/nature10983

19. Rhee J-K, Kim K, Chae H, Evans J, Yan P, Zhang B-T, et al. Integrated analysis of genome-wide DNA methylation and gene expression profiles in molecular subtypes of breast cancer. Nucleic Acids Research. 2013;41: 8464–8474. doi:10.1093/nar/gkt643

20. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature. 2012;490: 61–70. doi:10.1038/nature11412

21. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474: 609–615. doi:10.1038/nature10166

22. Vonlanthen J, Okoniewski MJ, Menigatti M, Cattaneo E, Pellegrini-Ochsner D, Haider R, et al. A comprehensive look at transcription factor gene expression changes in colorectal adenomas. BMC Cancer. 2014;14: 46. doi:10.1186/1471-2407-14-46

23. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, et al. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research. 2009;37: W202–8. doi:10.1093/nar/gkp335

24. Hannenhalli S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. Bioinformatics. 2008;24: 1325–1331. doi:10.1093/bioinformatics/btn198

25. Bulyk ML. Computational prediction of transcription-factor binding site locations. Genome Biology. 2003;5: 201. doi:10.1186/gb-2003-5-1-201

26. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, et al. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology. 2005;23: 137–144. doi:10.1038/nbt1053

27. Maher B. ENCODE: The human encyclopaedia. Nature. 2012;489: 46–48.

28. Sikora-Wohlfeld W, Ackermann M, Christodoulou EG, Singaravelu K, Beyer A. Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Computational Biology. 2013;9: e1003342. doi:10.1371/journal.pcbi.1003342

29. Brody LC, Biesecker BB. Breast cancer susceptibility genes. BRCA1 and BRCA2. Medicine (Baltimore). 1998;77: 208–226.

30. Fackenthal JD, Olopade OI. Breast cancer risk associated with BRCA1 and BRCA2 in diverse populations. Nature Reviews Cancer. 2007;7: 937–948. doi:10.1038/nrc2054

31. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Research. 2012;22: 1813–1831. doi:10.1101/gr.136184.111

32. Smyth GK. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical applications in genetics and molecular biology. 3. doi:10.2202/1544-6115.1027

33. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). OXFORD: Royal Statistical Society; 1995;57: 289–300.

34. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nature Reviews Genetics. 2009;10: 252–263. doi:10.1038/nrg2538

35. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;58: 267–288.

36. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of Statistical Software. 2010;33: 1–22.

37. Kim H, D'Andrea AD. Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway. Genes & Development. 2012;26: 1393–1408. doi:10.1101/gad.195248.112

38. Sharma P, Nag A. CUL4A ubiquitin ligase: a promising drug target for cancer and other human diseases. Open Biology. 2014;4: 130217. doi:10.1098/rsob.130217

39. García MJ, Saucedo-Cuevas LP, Muñoz-Repeto I, Fernández V, Robles MJ, Domingo S, et al. Analysis of DNA repair-related genes in breast cancer reveals CUL4A ubiquitin ligase as a novel biomarker of trabectedin response. Molecular Cancer Therapy. 2013;12: 530–541. doi:10.1158/1535-7163.MCT-12-0768

40. Underhill C, Toulmonde M, Bonnefoi H. A review of PARP inhibitors: from bench to bedside. Annals of Oncology. 2011;22: 268–279. doi:10.1093/annonc/mdq322

41. Moarii M, Boeva V, Vert J-P, Reyal F. Changes in correlation between promoter methylation and gene expression in cancer. BMC Genomics. 2015;16: 873. doi:10.1186/s12864-015-1994-2

42. Agrelo R, Cheng W-H, Setien F, Ropero S, Espada J, Fraga MF, et al. Epigenetic inactivation of the premature aging Werner syndrome gene in human cancer. Proceedings of the National Academy of Sciences. 2006;103: 8822–8827. doi:10.1073/pnas.0600645103

43. Esteller M. Epigenetic gene silencing in cancer: the DNA hypermethylome. Human Molecular Genetics. 2007;16 Spec No 1: R50–9. doi:10.1093/hmg/ddm018

44. Zona S, Bella L, Burton MJ, Nestal de Moraes G, Lam EW-F. FOXM1: an emerging master regulator of DNA damage response and genotoxic agent resistance. Biochim Biophys Acta. 2014;1839: 1316–1322. doi:10.1016/j.bbagrm.2014.09.016

45. Pilarsky C, Wenzig M, Specht T, Saeger HD, Grützmann R. Identification and validation of commonly overexpressed genes in solid tumors by comparison of microarray data. Neoplasia. 2004;6: 744–750. doi:10.1593/neo.04277

46. Bektas N, Haaf AT, Veeck J, Wild PJ, Lüscher-Firzlaff J, Hartmann A, et al. Tight correlation between expression of the Forkhead transcription factor FOXM1 and HER2 in human breast cancer. BMC Cancer. 2008;8: 42. doi:10.1186/1471-2407-8-42

47. Teh M-T, Gemenetzidis E, Chaplin T, Young BD, Philpott MP. Upregulation of FOXM1 induces genomic instability in human epidermal keratinocytes. Molecular Cancer. 2010;9: 45. doi:10.1186/1476-4598-9-45

48. Priller M, Pöschl J, Abrão L, Bueren von AO, Cho Y-J, Rutkowski S, et al. Expression of FoxM1 is required for the proliferation of medulloblastoma cells and indicates worse survival of patients. Clinical Cancer Research. 2011;17: 6791–6801. doi:10.1158/1078-0432.CCR-11-1214

49. Chu X-Y, Zhu Z-M, Chen L-B, Wang J-H, Su Q-S, Yang J-R, et al. FOXM1 expression correlates with tumor invasion and a poor prognosis of colorectal cancer. Acta Histochem. 2012;114: 755–762. doi:10.1016/j.acthis.2012.01.002

50. Ren B, Cam H, Takahashi Y, Volkert T, Terragni J, Young RA, et al. E2F integrates cell cycle progression with DNA repair, replication, and G(2)/M checkpoints. Genes & Development. 2002;16: 245–256. doi:10.1101/gad.949802

51. Biswas AK, Johnson DG. Transcriptional and nontranscriptional functions of E2F1 in response to DNA damage. Cancer Research. 2012;72: 13–17. doi:10.1158/0008-5472.CAN-11-2196

52. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA. Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets. Genome Medicine. 2012;4: 41. doi:10.1186/gm340

53. Kachhap SK, Rosmus N, Collis SJ, Kortenhorst MSQ, Wissing MD, Hedayati M, et al. Downregulation of homologous recombination DNA repair genes by HDAC inhibition in prostate cancer is mediated through the E2F1 transcription factor. PLoS ONE. 2010;5: e11208. doi:10.1371/journal.pone.0011208

54. Zaika E, Wei J, Yin D, Andl C, Moll U, El-Rifai W, et al. p73 protein regulates DNA damage repair. FASEB Journal. 2011;25: 4406–4414. doi:10.1096/fj.11-192815

55. Zhan M, Riordon DR, Yan B, Tarasova YS, Bruweleit S, Tarasov KV, et al. The B-MYB transcriptional network guides cell cycle progression and fate decisions to sustain self-renewal and the identity of pluripotent stem cells. PLoS ONE. 2012;7: e42350. doi:10.1371/journal.pone.0042350

56. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

57. Maienschein-Cline M, Zhou J, White KP, Sciammas R, Dinner AR. Discovering transcription factor regulatory targets using gene expression and binding data. Bioinformatics. 2012;28: 206–213. doi:10.1093/bioinformatics/btr628

58. de Gramont A, Watson S, Ellis LM, Rodón J, Tabernero J, de Gramont A, et al. Pragmatic issues in biomarker evaluation for targeted therapies in cancer. Nature Reviews Clinical Oncology. 2015;12: 197–212. doi:10.1038/nrclinonc.2014.202

59. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to

uncover genotype-phenotype interactions. Nature Reviews Genetics. 2015;16: 85–97. doi:10.1038/nrg3868

60. Kristensen VN, Lingjaerde OC, Russnes HG, Vollan HKM, Frigessi A, Børresen-Dale A-L. Principles and methods of integrative genomic analyses in cancer. Nature Reviews Cancer. 2014;14: 299–313. doi:10.1038/nrc3721

61. Li Y, Liang M, Zhang Z. Regression analysis of combined gene expression regulation in acute myeloid leukemia. PLoS Computational Biology. 2014;10: e1003908. doi:10.1371/journal.pcbi.1003908

62. Setty M, Helmy K, Khan AA, Silber J, Arvey A, Neezen F, et al. Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma. Molecular Systems Biology. 2012;8: 605. doi:10.1038/msb.2012.37

63. Jiang P, Freedman ML, Liu JS, Liu XS. Inference of transcriptional regulation in cancers. Proceedings of the National Academy of Sciences. 2015;112: 7731–7736. doi:10.1073/pnas.1424272112

64. Liu Q, Su P-F, Zhao S, Shyr Y. Transcriptome-wide signatures of tumor stage in kidney renal clear cell carcinoma: connecting copy number variation, methylation and transcription factor activity. Genome Medicine. 2014;6: 117. doi:10.1186/s13073-014-0117-z

65. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, et al. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Research. 2000;28: 316–319.

66. Inui M, Martello G, Piccolo S. MicroRNA control of signal transduction. Nature Reviews Molecular Cell Biology. 2010;11: 252–263. doi:10.1038/nrm2868

Chapter Five: General Discussion

Genomic instability is an *enabling hallmark* of cancer, which generates genetic diversity that fosters the acquisition of other cancer hallmarks (1). In order to prevent genomic instability, cells have evolved sophisticated DNA repair machinery to detect, signal and repair a diverse array of genotoxic lesions. Hundreds of DNA repair genes have been identified as components of this machinery. From a genetic perspective, the behaviour of DNA repair genes is much like that of tumour suppressor genes in a sense that defects in both types of genes predispose to cancer development (1, 2). This highlights the importance of studying DNA repair in order to understand cancer etiology.

Studying DNA repair is also critical in cancer management. The majority of the DNA repair genes can be functionally grouped into five major DNA repair pathways, each specific for the repair of one type or limited types of DNA lesions (described in Chapter One and Two). It is now clear that upregulated DNA repair pathways can enable tumour cells to survive damages induced by chemotherapy and/or radiotherapy, while downregulated DNA repair pathways render tumours sensitive to these DNA-damaging therapies (3, 4). Accordingly, small-molecule inhibitors that directly target different repair pathways are being developed (3, 5), and it is very likely that in the near future these inhibitors can be used in combination with distinct chemotherapeutic agents and/or radiation to improve treatment efficacy. Furthermore, defects in a DNA repair pathway that arise during tumour initiation and progression often make tumours become "addicted" to another compensatory repair pathway for survival. This provides a vulnerability that can be therapeutically exploited using the principle of SL (described in Chapter One and Two) for the development of a novel class of targeted therapies (6). In fact *olaparib*, the first drug regime of this class, has recently been approved by the US Food and Drug Administration (FDA) for remedying *BRCA*-mutated, heavily pretreated ovarian tumours; and efforts are underway to expand the use of *olaparib* in other cancer types (5, 6).

Taken together, these developments suggest that we are now entering a new era of cancer treatment, in which optimised therapy can be tailored for individual cancer patients based on the DNA repair status of the tumour, rather than on its histologic appearance and/or tissue of origin (3, 5, 7, 8). In light of the precision medicine initiative recently announced by the US government (9), the importance of a thorough understanding of the DNA repair system for better cancer therapy can hardly be overemphasised.

In the last decade, enormous progress has been made towards characterising the constituent gene components for each repair pathway and understanding how the gene-specific genetic defects

affect cancer predisposition and/or treatment response. However, outstanding challenges still remain before DNA repair knowledge can be successfully applied for precision medicine. The studies presented in this thesis attempt to address some of these fundamental challenges through a computational analysis of DNA repair pathways in breast cancer. This chapter summarises the key outcomes of the thesis and discusses possible future directions.

5.1 A High-quality manual-curation of the repair pathways

One challenge facing us is to provide an accurate, comprehensive and up-to-date annotation for each repair pathway. The quality of such work is of fundamental importance for applying any systems biology approaches, including pathway and network analyses, to the study of DNA repair. Much effort has been invested to construct prime pathway databases via manual curation (10, 11); however, at the time I commenced this thesis (2012), these common pathway resources failed to keep their DNA repair-related content up-to-date. For example, the Reactome database (10) had not updated its DNA repair entries since 2003, even though dozens of new DNA repair genes had been identified after that time. A similar situation was seen with the KEGG database (11). The first project of this thesis (Chapter Two) was therefore committed to performing a high-quality manual curation for each repair pathway.

This curation process involved extracting relevant knowledge from literature and consulting a domain expert. The components in each curated pathway include not only enzymes such as ATPases, polymerases and DNA ligases, that directly conduct the repair function, but also various histone modification and chromatin remodelling factors that enable access to DNA lesions. Moreover, various posttranslational modification factors, including those involved in methylation, phosphorylation, ubiquitination, sumoylation and acetylation are also incorporated as they have a substantial impact on the repair pathway activity. After the curation, all the knowledge about a repair pathway, which was widely scattered over a range of knowledge domains, was assembled into a pathway diagram using Cytoscape (12). And for each reaction included in the pathway diagrams, a detailed description is given that enables mechanistic understanding of the pathway content. This curation work provides a solid foundation for the computational analysis presented in the following chapters.

Currently, our understanding of the DNA repair pathways continues to expand. In addition to cataloguing new DNA repair genes, in the future we will also need to annotate exact transcripts that function in a given repair pathway and/or a given tissue. This can be a critical issue for pathway analysis as more than 90% of the human genome is subjected to alternative splicing and multiple transcripts from the same gene may have different or even opposing functions (13, 14).

5.2 Quantifying DNA repair pathway dysregulation at tumour level

Another challenge lies in how to accurately and efficiently evaluate the functional status of each repair pathway at the level of individual patients. This issue is of key importance for applying DNA repair knowledge for precision medicine, as described at the beginning of this Chapter. To date, pathway analysis has become a common practice in biomedical research; however, most of these analyses capture only pathways that differ between two phenotypic conditions, such as disease and normal, and hence are not able to provide patient-specific pathway aberrance information.

In Chapter Three, I performed a personalised pathway analysis independently for four large breast cancer cohorts (about 3,000 tumours in total) to investigate the status of HR pathway dysregulation in individual sporadic breast tumours, its association with HR repair deficiency and its impact on tumour characteristics. Specifically, I calculated an *HR* score for each breast tumour to quantify the extent of HR pathway dysregulation in that tumour. Based on the score, I found HR dysregulation is prevalent in breast tumours but the extent differs greatly between and within the previously well-recognised breast cancer subtypes, underscoring the necessity of personalised analysis. Furthermore, I found that HR pathway dysregulation reflects HR repair deficiency, suggesting that the *HR* score can be used as a convenient way for detecting HR repair deficiency in individual tumours. Most importantly, I uncovered a novel association between HR pathway dysregulation and CIN in sporadic breast cancer. Although the importance of CIN in tumour evolution and drug resistance has been highlighted in recent extensive studies, the molecular basis of CIN in sporadic cancers remains poorly understood. The novel association revealed here indicates that dysregulated HR may be an important contributor to CIN in sporadic breast cancer, and thus facilitates future experiments to pinpoint the causative factors of CIN in sporadic breast cancer as well as in other sporadic cancers.

Variability of HR dysregulation can also exist within individual tumours. A bulk tumour sample is usually composed of thousands of tumour cells that belong to genetically distinct subclonal populations. The relative abundances of these populations, known as the subclonal structure of a tumour, vary dynamically in different micro-environments or after exposure to cancer drugs, enabling the tumour to obtain metastatic potential or survive therapy (15-17). It is possible that the cellular status of HR, along with the status of other DNA repair pathways, may substantially affect how the subclonal tumour cell populations acquire new genetic alterations, evolve, and compete with each other; this topic, however, has rarely been touched upon so far due to a lack of study tools. With the rapid advance of single-cell sequencing technologies in recent years and the ever-falling cost of sequencing, it is becoming feasible to obtain genomic datasets for

different parts of a tumour. As long as there are enough samples being measured, Pathifier can be applied directly to such datasets to interrogate pathway variation within tumours.

5.3 Integrating multi-omics data to decipher mechanisms of DNA repair dysregulation

A third challenge is how to integrate multi-omics data to facilitate a mechanistic understanding of DNA repair dysregulation in cancer. In Chapter Three, I demonstrated that HR pathway dysregulation is prevalent in breast cancer. In earlier studies, numerous research groups documented the aberrant expression of individual DNA repair genes in various cancer types. The transcriptional changes of DNA repair genes and/or pathways in cancer provide the first line of evidence for the altered DNA repair status in tumour, and hence have important implications for biomarker discovery and treatment selection. However, although large-scale genomic projects, such as TCGA, have generated a wealth of genomic data at multiple molecular levels, it remains challenging to combine the information from these different levels of data to provide insights into the underlying biology of DNA repair dysregulation in cancer.

In Chapter Four, I systematically dissected the contributions of DNA copy number alteration (CNA), DNA methylation at gene promoter regions (DM) and expression changes of transcriptional factors (TFs) to the differential expression of individual DNA repair genes in breast tumour versus normal samples; these gene-specific results were summarised at pathway level to examine whether different DNA repair pathways are affected in distinct manner. In particular, I developed a regularised linear regression-based statistical framework to identify relevant TFs for each DNA repair gene from a comprehensive list of 1391 manually curated human TFs (18). This framework takes into account the contributions of genetic and epigenetic changes to gene expression variations, and overcomes the instability inherent to the regularised linear regression methods. The results suggest that CNA and expression changes of TFs are major causes of DNA repair dysregulation in breast cancer, and that a subset of the identified TFs may have global impact on the dysregulation of different repair pathways. The work presented in this chapter provides novel insights into DNA repair dysregulation in breast cancer. These insights improve our understanding of the molecular basis of the DNA repair biomarkers identified thus far, and have the potential to inform future biomarker discovery.

Integrating multi-omics data for better understanding of biological systems is still in its infancy, and much still remain unknown regarding how gene expression is regulated in normal cells, and how this exquisite control is compromised in cancer as well as in other diseases (19). In the future, novel methods taking advantage of the ever-increasing multi-omics data, especially chromatin immunoprecipitation sequencing (ChIP-Seq) data, would be the key to obtain a more comprehensive understanding of DNA repair dysregulation in cancer.

5.4 Concluding remarks

Cancer is an increasingly serious concern, especially with an aging population. A thorough understanding of the DNA repair system will ultimately lead to tremendous improvements in cancer treatment. The three pieces of work presented in this thesis form a coherent research story, starting from building a knowledge base for DNA repair (Chapter Two), followed by personalised analysis of DNA repair pathway dysregulation (Chapter Three), then ending up with elucidating underlying molecular mechanism of DNA repair dysregulation (Chapter Four). The results generated from these studies deepen our understanding of the complex DNA repair system and also inform future studies for better cancer therapy.

There are two emerging directions in the field of DNA repair research, which were not touched upon in this thesis due to limitations of time and resources, but are worth exploring in the future. First, although DNA repair pathways are generally considered as mutually exclusive mechanisms responsible for distinct types of damages, a more accurate view is that they function in a dynamic and interconnected network. Recent studies (20-23) provide strong evidence that "crosstalks" exist between different DNA repair mechanisms; namely, these repair pathways not only share common proteins, but also operate synergistically to repair lesions. Moreover, proteins that are involved in multiple repair pathways, termed DNA repair hubs, are frequently compromised in various cancers, and have been proposed as potential targets for targeted therapy (24).

In a broader view, DNA repair pathways are linked and coordinated with other important cellular pathways, such as cell cycle checkpoint and apoptosis, all of which can be included in the intrinsic network responsible for maintaining genomic instability. In light of this view, it is interesting to see that, among 186 KEGG pathways (11) and 674 Reactome pathways (10), the pathway dysregulation score of the cell cycle pathway is most highly correlated with the *HR* score; and the dysregulation score of DNA repair pathways other than HR also have much higher correlation with the *HR* score compared with most of other pathways (Chapter Three, Supplementary Table S3 and S4; the pathway dysregulation score and *HR* score were calculated in the same way). In the future, more systems-level analyses are needed to further characterise this genome-maintaining network, of which the results will be of great importance for dissecting the complexity of the cellular DNA repair system.

Second, although it is a well-known issue, I still would like to emphasise the importance of applying proteomics techniques for DNA repair research. Gry et al. (25) compared the RNA profiles and protein profiles of 1066 genes in 23 human cell lines, and found that the mean correlations between protein and RNA levels was only about 0.22. Besides, recent studies (26-28) have shown that a variety of posttranslational modifications have a substantial impact on various

aspects of DNA repair, including damage recognition, signal transduction and loading of the repair proteins. With the mass spectrometry-based proteomics technologies becoming more mature (29), these new technologies will play increasingly important roles in our understanding of the DNA repair system.

In conclusion, we will very likely see the realisation of precision medicine in the next decade or two, which will improve human health and life expectancy in an unprecedented manner. With the advances of the various genomic profiling techniques and the advent of novel pathway and network analysis approaches, I hope that systems biology studies in the field of DNA repair, including those herein, will contribute to the early achievement of this great goal.

References

1. Hanahan,D. and Weinberg,R.A. (2011) Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674.

2. Kinzler,K.W. and Vogelstein,B. (1997) Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature*, **386**, 761–763.

3. Curtin,N.J. (2012) DNA repair dysregulation from cancer driver to therapeutic target. *Nature Reviews Cancer*, **12**, 801–817.

4. Helleday,T., Petermann,E., Lundin,C., Hodgson,B. and Sharma,R.A. (2008) DNA repair pathways as targets for cancer therapy. *Nature Reviews Cancer*, **8**, 193–204.

5. Pearl,L.H., Schierz,A.C., Ward,S.E., Al-Lazikani,B. and Pearl,F.M.G. (2015) Therapeutic opportunities within the DNA damage response. *Nature Reviews Cancer*, **15**, 166–180.

6. Lord,C.J., Tutt,A.N.J. and Ashworth,A. (2015) Synthetic lethality and cancer therapy: lessons learned from the development of PARP inhibitors. *Annual Review of Medicine*, **66**, 455–470.

7. Goldstein,M. and Kastan,M.B. (2015) The DNA damage response: implications for tumor responses to radiation and chemotherapy. *Annual Review of Medicine*, **66**, 129–143.

8. Torgovnick,A. and Schumacher,B. (2015) DNA repair mechanisms in cancer development and therapy. *Frontiers in Genetics*, **6**, 157.

9. Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *The New England Journal of Medicine*, **372**, 793–795.

10. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B., *et al.* (2010) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research*, **39**, D691–D697.

11. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, **28**, 27–30.

12. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models

of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

13. Wang,E.T., Sandberg,R., Luo,S., Khrebtukova,I., Zhang,L., Mayr,C., Kingsmore,S.F., Schroth,G.P. and Burge,C.B. (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

14. Khatri,P., Sirota,M. and Butte,A.J. (2012) Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology*, **8**, e1002375.

15. Gawad,C., Koh,W. and Quake,S.R. (2016) Single-cell genome sequencing: current state of the science. *Nature Reviews Genetics*, **17**, 175–188.

16. Marusyk,A., Almendro,V. and Polyak,K. (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nature Reviews Cancer*, **12**, 323–334.

17. Burrell,R.A., McGranahan,N., Bartek,J. and Swanton,C. (2013) The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*, **501**, 338–345.

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

19. Handley,A., Schauer,T., Ladurner,A.G. and Margulies,C.E. (2015) Designing Cell-Type-Specific Genome-wide Experiments. *Molecular Cell*, **58**, 621–631.

20. Peng,M., Xie,J., Ucher,A., Stavnezer,J. and Cantor,S.B. (2014) Crosstalk between BRCA-Fanconi anemia and mismatch repair pathways prevents MSH2-dependent aberrant DNA damage responses. *The EMBO Journal*, **33**, 1698–1712.

21. Mouw,K.W. and D'Andrea,A.D. (2014) Crosstalk between the nucleotide excision repair and Fanconi anemia/BRCA pathways. *DNA Repair*, **19**, 130–134.

22. Haynes,B., Saadat,N., Myung,B. and Shekhar,M.P.V. (2015) Crosstalk between translesion synthesis, Fanconi anemia network, and homologous recombination repair pathways in interstrand DNA crosslink repair and development of chemoresistance. *Mutation Research-Reviews in Mutation Research*, **763**, 258–266.

23. Herrero,A.B., Martín-Castellanos,C., Marco,E., Gago,F. and Moreno,S. (2006) Cross-talk between nucleotide excision and homologous recombination DNA repair pathways in the mechanism of action of antitumor trabectedin. *Cancer Research*, **66**, 8155–8162.

24. Dietlein,F., Thelen,L. and Reinhardt,H.C. (2014) Cancer-specific defects in DNA repair pathways as targets for personalized therapeutic approaches. *Trends in Genetics*, **30**, 326–339.

25. Gry,M., Rimini,R., Strömberg,S., Asplund,A., Pontén,F., Uhlén,M. and Nilsson,P. (2009) Correlations between RNA and protein expression profiles in 23 human cell lines. *BMC Genomics*, **10**, 365.

26. Peng,G. and Lin,S.-Y. (2011) Exploiting the homologous recombination DNA repair network for targeted cancer therapy. *The World Journal of Clinical Oncology*, **2**, 73–79.

27. Ulrich,H.D. (2012) Ubiquitin and SUMO in DNA repair at a glance. *Journal of Cell Science*, **125**, 249–254.

28. Dou,H., Huang,C., Van Nguyen,T., Lu,L.-S. and Yeh,E.T.H. (2011) SUMOylation and de-SUMOylation in response to DNA damage. *FEBS Letters*, **585**, 2891–2896.

29. Gstaiger,M. and Aebersold,R. (2009) Applying mass spectrometry-based proteomics to genetics, genomics and network biology. *Nature Reviews Genetics*, **10**, 617–627.

Appendix 1: Unpublished results from the curation work

A1.1 Homologous recombination

Of all DNA lesions, double-strand breaks (DSBs) are generally considered to be the most toxic and can be lethal if left unrepaired. Homologous Recombination (HR) and Non-homologous End-joining (NHEJ) are the two main repair pathways for DSBs, the choice of which depends on cell cycle stage and the structure of the DNA ends. HR occurs during the S and G2 phases of the cell cycle as it needs a homologous sister chromatid to restore the lost DNA information at the break site, resulting in a repair that is more accurate than with other mechanisms.

HR is proposed to occur in several steps. First, the damaged DNA is recognized by the MRE11-RAD50-NBS1 (MRN) complex, which also plays a role in the recruitment and activation of the Ataxia Telangiectasia Mutated (ATM) kinase. Activated ATM initiates a signaling cascade that results in the phosphorylation of many substrates, including histone variant H2AX. Phosphorylated H2AX (γH2AX) in turn recruits MDC1 to sites of DNA damage, which acts as an adaptor protein that help many other DDR proteins to be targeted to the damage sites, such as RNF8, RNF168, BRCA1 and 53BP1.

At the same time as these repair proteins are recruited, the DSB ends are undergoing a two-phase resection, with limited resection initiated by MRN and CtIP, and extensive resection mediated by the BLM helicase and DNA2 exonuclease, or the EXO1 exonuclease alone. The replication protein A (RPA) recognizes and binds the 3' single-stranded (ssDNA) tail generated by the resection, and is then replaced by RAD51 with the assistance of other factors, including BRCA2. The RAD51-bound ssDNA (termed RAD51 nucleoprotein filament) searches for and invades homologous template. The invading strand is extended by synthesis of new DNA, and can facilitate the exchange of the homologous DNA strands to generate a D loop or a Holliday junction structure, which can then be resolved by various enzymes, respectively, to complete the repair.

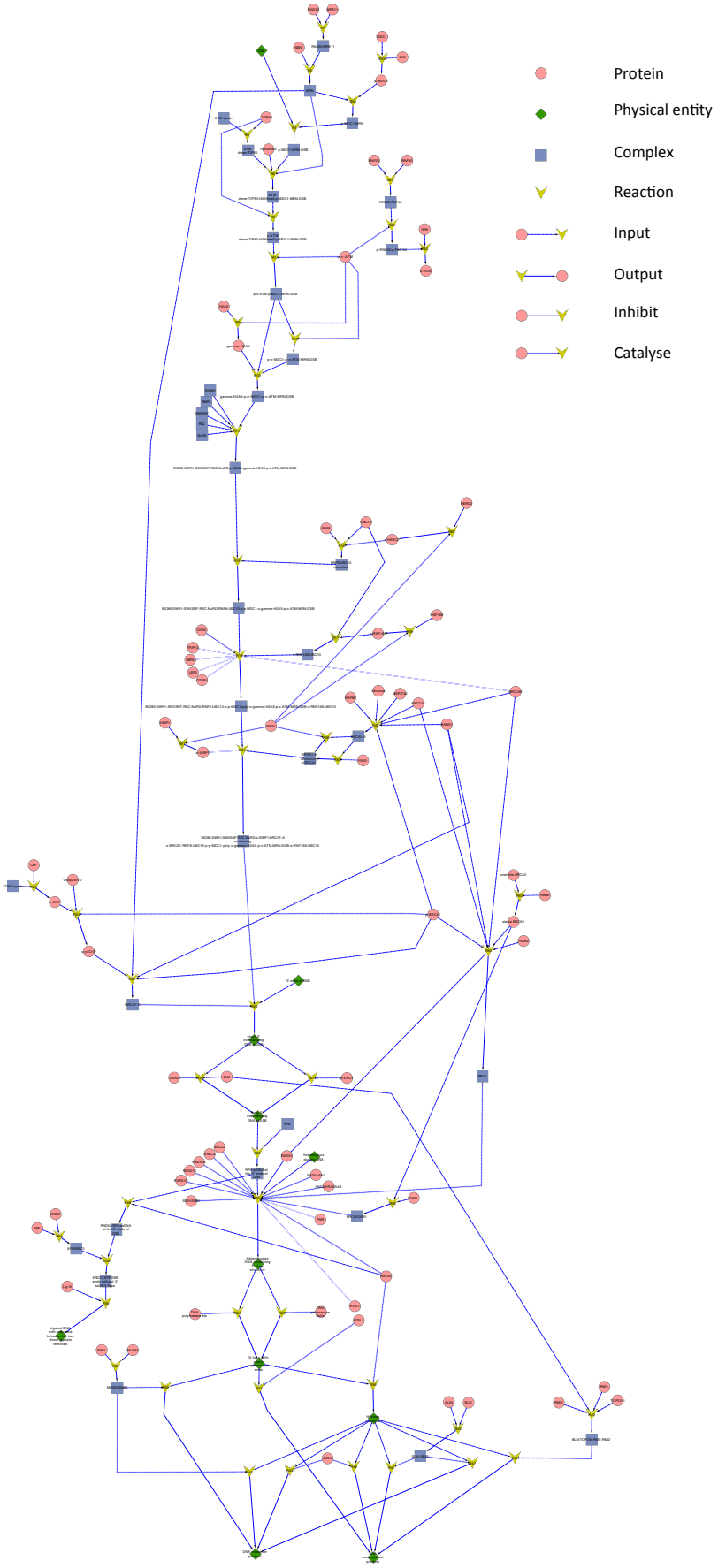


Figure 1 Homologous recombination (HR) pathway figure. The original PDF version is deposited at UQ eSpace.

A1.1.1 Reaction Description

R1 Formation of RAD50:MRE11 Complex

See description for R3.

R2 Association of RAD50:MRE11 with NBS1 via MRE11 interaction

See description for R3.

R3 Association of MRN with sites of DSB

The MRN complex is an assembly of two MRE11 subunits, two RAD50 subunits, and one NBS1 subunit. MRE11 has ssDNA endonuclease and 3' to 5' exonuclease activities important for the initial step of DNA end resection in HR; RAD50 associates with DSB ends and interacts with MRE11; NBS1 also interacts with MRE11 and it has been shown to recruit ATM to a DSB through its C-terminus and interacts with MDC1 via its N-terminus.

MRN complex has multiple roles in DSB repair – it recognizes the broken ends, acts as a DNA-bridging scaffold to prevent chromosome separation, catalyzes the activation of ATM in conjunction with other proteins (such as TIP60), and participates in the early steps of end resection at DSBs. By promoting the activation of the ATM kinase, this complex is also involved in setting up a cell cycle checkpoint response in reaction to DNA damage [1, 2].

R4 formation of ATM-TIP60 Complex

See description for R6.

R5 Recruitment of ATM-TIP60 Complex by MRN, and activation of TIP60 histone acetyltransferase activity by Binding to H3K9me3

See description for R6.

R6 Acetylation of ATM by TIP60

ATM and tumor suppressor TIP60 form a stable complex that is then recruited to DSBs by MRN. Following the recruitment, TIP60 binds to a histone variant H3 tri-methylated on lysine 9 (H3K9me3), which activates the histone acetyltransferase (HAT) activity of Tip60, leading to the subsequent acetylation and activation of the kinase activity of ATM [3-5].

R7 Intermolecular autophosphorylation and dissociation of dimeric ATM Complexes

ATM normally remains inactive in human cells as a dimer in which it is unable to phosphorylate other cellular substrates. In response to DNA damage, and with the presence of MRN in the damage site, the kinase domain of one ATM molecule phosphorylates Ser-1981 of another ATM molecule in the same dimer, and the phosphorylated ATM is then dissociated from the complex with its phosphorylation activity activated [3, 6, 7]. It has been recently shown that the activation process of ATM also involves autophosphorylation on Ser-367, Ser-1893, Ser-2996 and acetylation on Lys-3016 [8].

R9 Phosphorylation of histone H2AX at Serine-139 by ATM at the site of DSB

Immediately following recruitment of ATM to DSB sites, it phosphorylates histone variant H2AX on Ser139, producing γH2AX that is required for DNA damage signal amplification and subsequent accumulation of numerous DNA damage response (DDR) proteins at DSBs sites [3, 6, 9].

In addition to ATM, recent research shows H2AX can also be phosphorylated by ATR and DNA-PK in response to DNA damage. These two proteins are members of the phosphoinositide 3-kinase related protein kinase (PIKK) family that also contains ATM. ATM and DNA-PKcs display functional redundancy in phosphorylating H2AX following ionizing radiation, while ATR is more important for H2AX phosphorylation in response to DNA damage that would slow or stall replication forks [9].

R10 Association of γH2AX and phosphorylated MDC1

MDC1 directly binds γH2AX for its recruitment to DSB sites. After its recruitment, MDC1 facilitate more ATM and MRN recruitment to further promote γH2AX spreading for distances up to 1-2 megabases around DSBs. It has been suggested that the recruitment of γH2AX and the phosphorylation of MDC1 at DSB sites provides a docking site for many components of HR pathway [10, 11].

R11 Recruitment of chromatin-remodeling complex INO80, SWR1, SWI/SNF, RSC and NuRD

Chromatin-remodeling complexes, such as INO80, SWI/SNF, RSC and NuRD, are thought to be involved directly in HR repair. In particular, recent experiments suggested that INO80, which is recruited in a γH2AX-dependent manner, promotes histone eviction around the DSB site, and stimulates ssDNA formation and checkpoint activation; SWI/SNF, which is also recruited in a γH2AX-dependent manner, can further promote H2AX phosphorylation to form a positive feedback loop; RSC is recruited to DSB lesions very early and mobilizes nucleosomes to promote loading of MRN complex; NuRD facilitates RNF8/RNF168-dependent histone ubiquitination to facilitate ubiquitin-dependent retention of RNF168 and BRCA1[12, 13].

R13 Phosphorylation of MDC1 by ATM

Immediately following phosphorylation of H2AX by ATM, MDC1 is also phosphorylated by ATM and then recruited to DSB sites via binding to γH2AX. Phosphorylated form of MDC1 could serve as a docking site for the FHA domain of RNF8 [3, 6].

R14 Formation of RNF8-UBC13 complex

The E3 ubiquitin ligase RNF8 forms a complex with E2 ubiquitin-conjugating enzyme UBC13 to ubiquitinate γH2AX near the DSB sites [14]. The formation of this complex is facilitated by HERC2, another E3 ubiquitin ligase that interacts with the FHA domain of RNF8, and stimulates the ubiquitin ligase activity of RNF8 [15].

R15 Recruitment of RNF8-UBC13 complex which then mono-ubiquitinate gamma H2AX

The phosphorylated MDC1 serves as a docking site for the recruitment of RNF8 and UBC13 complex, which mediates the mono-ubiquitination of γH2AX and provides a docking site for the recruitment of RNF168 [12].

R17 Formation of RNF168-UBC13 complex

See description for R18.

R18 Recruitment of RNF168-UBC13 complex which then poly-ubiquitinate gamma H2AX

The E3 ubiquitin ligase RNF168 forms a complex with E2 ubiquitin-conjugating enzyme UBC13, which binds and amplifies the RNF8-mediated ubiquitination of γH2AX, generating ubiquitin chains that are required for the accumulation and retention of 53BP1 and BRCA1-A complex [16, 17]. The chromodomain helicase DNA-binding protein 4 (CHD4), which is a chromatin remodelling factor, may also play an essential role in this process as depletion of CHD4 disrupts the recruitment of RNF168 [18].

On the other hand, as uncontrolled amplification of chromatin ubiquitination could have deleterious consequences, there are at least five deubiquitination enzymes (BRCC36, USP3, OTUB1, TRIP12, UBR5) that can counteract the activity of RNF168, and therefore confine the ubiquitination to DNA lesions [19, 20].

R20 The formation of BRCA1-A Complex (BARD1 - BRCA1- abraxas - RAP80 - BRCC36 - BRCC45 - MERIT40)

The BRCA1-A complex consists of BRCA1-BARD1 heterodimer, RAP80 (also known as UIMC1), Abraxas (also known as FAM175A or CCDC98), BRCC36 (also known as BRCC3), BRCC45 (also known as BRE), RAP80 (also known as UIMC1), NBA1 (also known as MERIT40). BRCC36 and BRCC45 are thought to facilitate the E3 ubiquitin ligase activity of the BRCA1–BARD1 heterodimer; NBA1 stabilizes various components of the BRCA1-A complex; Abraxas-RAP80 sub-complex specifically binds to polyubiquitin chains present on γH2AX at DNA damage sites [21-23].

R21 Recruitment of BRCA1-A complex

The BRCA1-A complex directly binds to the polyubiquitin chains generated by RNF8 and RNF168 through the Abraxas-RAP80 sub-complex. It has been suggested that this complex may functions to control BRCA1 activity for DSB end resection and prevent excess HR activity by suppressing BRCA1-C complex [21-23]. 53BP1 binds to methylated histones, which could possibly only be recognized after RNF8–RNF168–UBC13-mediated polyubiquitylation. This protein is critical for the control of DSB repair, promoting NHEJ and inhibiting the 5′ end resection needed for HR. Loss of 53BP1 partially rescues the HR defect of BRCA1 mutant cells [24, 25].

R23 Phosphorylation of CtIP at Thr-847 and Ser-327 by CDK2-cycline A Complex

CtIP (also known as RBBP8) acts together with BRCA1 and MRN complex to promote initial end resection for DSBs, which is a critical step for HR repair. Actually, HR capacity is maximized in S and G2 through phosphorylation of CtIP on Thr-847 and Ser-327 by CDK2-cycline A complex [26]. As DSB end resection is essential for HR but not for NHEJ, this modification of CtIP many represent a mechanism by which human cells modulate DSB repair pathway choice during cell cycle [27].

R24 Ubiquitination of CtIP by BRCA1

In S and G2, CDK-dependent phosphorylation of CtIP causes its interaction with the BRCT domains of BRCA1, resulting in the activation of BRCA1 E3 ligase activity and subsequent CtIP ubiquitination [22, 28].

R25 The formation of BARD1-BRCA1-CtIP - MRN Complex (BRCA1-C Complex)

BRCA1, CtIP, and MRN form a complex when cells enter the S and G2 phases of the cell cycle. It has been proposed that phosphorylated CtIP promotes DNA end resection by interacting and stimulating the nuclease activity of the MRN complex [22, 29].

R26 Initial DSB end resection by BRCA1-C complex

The resection of DSBs to generate ssDNA tails is a two-step process, which occurs after the accumulation of a number of upstream HR proteins, such as MDC1, RNF8 and BRCA1. The initial DSB end resection by BRCA1-C complex is likely to be conducted mainly by the nuclease activity of the MRN complex. CtIP promotes this resection by interacting with MRN and stimulating its nuclease activity. BRCA1 may function as a scaffold to stabilize MRN and CtIP as it is not known to have any nuclease or helicase domains [22, 30, 31].

R27 Extensive resection by EXO1

Following the initial resection of DSB ends, a long-range resection is conducted by either DNA Exonuclease I (EXO1) or by Bloom Syndrome Protein (BLM) and DNA replication helicase 2 (DNA2). This process generates ssDNA tails at the break site, which is essential for the subsequent cell cycle checkpoints activation and for RAD51 mediated strand invasion into

homologous duplex DNA. EXO1 serves as a resection nuclease and plays a predominant role in this process [30, 32, 33].

R28 Extensive resection by DNA2 and BLM

BLM and DNA2 form a complex and interact functionally to conduct extensive resection of DSB ends [34]. This long-range resection machine is redundant with the one involving EXO1, and it remains unknown what determines the choice between these two machines [30, 35].

R29 Association of RPA Complex with ssDNA

Replication protein A (RPA) is an ssDNA binding protein that has three subunits (RPA1, RPA2 and RPA3), which is recruited to DSB sites following the formation of 3' ssDNA by DSB end resection. It plays a critical role in stabilizing the 3' ssDNA regions, and therefore is essential for the assembly of RAD51 filaments on RPA-coated ssDNA and activation of the cell cycle check point by ATR [36-38].

R30 The formation of BARD1- BRCA1 - PALB2 - BRCA2 - RAD51 - BRCC36 - BRCC45 Complex (BRCC Complex)

BRCC is an ubiquitin E3 ligase complex consisting of BRCA1, BARD1, BRCA2 (also known as FANCD1), PALB2 (also known as FANCN), RAD51, BRCC36 and BRCC45. In this complex, BRCC36 and BRCC45 have been found to promote the E3 ubiquitin ligase activity of the BRCA1–BARD1 heterodimer. PALB2, a partner and localizer of BRCA2, has been shown to serve as the molecular adaptor between BRCA1 and BRCA2. BRCA1 is thought to fine tune HR partly through its modulatory role in the PALB2-dependent loading of BRCA2-RAD51 repair machinery at DNA breaks [21, 22, 39].

R31 Formation of RAD51-ssDNA nucleoprotein filament and DNA D-loop structure

One of the key steps in HR is the loading of RAD51, an evolutionarily conserved recombinase, onto ssDNA to form a nucleoprotein filament. This filament is the catalyst for strand invasion into homologous duplex DNA, resulting in the formation of a D loop structure. The above process requires the participation of the tumor suppressor BRCA2, which acts as a mediator binding both ssDNA and RAD51. This process also depends on the concert action of a number of partner proteins, including the five RAD51 paralogs (RAD51B, RAD51C, RAD51D, XRCC2 and XRCC3)

[40], RAD52 [41], RAD54 and its paralog RAD54B [42, 43], RAD51AP1 [44, 45], and the two ssDNA binding proteins SSB1 and SSB2 [46, 47].

Apart from the positive role that HR has in maintaining genome stability, inappropriate hypercombination can also cause genomic instability and cancer. Proteins that restrict unscheduled HR include RETL1, which interferes with the formation of RAD51-ssDNA filament [48], and PCNA-associated recombination inhibitor (PARI), which promotes the disassembly of the D loop structure [49].

R34 Extension of 3' invading strand by DNA polymerase eta

See description for R35.

R35 Extension of 3' invading strand by DNA polymerase delta

After resection to create single-stranded 3′ overhangs, strand invasion allows for 3′ extension. Branch migration of the resulting Holliday Junction allows for release of the invading strand, which subsequently anneals to the opposite side of the original break. Recent studies showed that DNA polymerase eta [50, 51] or more likely DNA polymerase delta [52, 53] can perform 3′ end extension at a D-loop.

R36 Holliday junction formation mediated by RAD52

Capture of the second DSB end by annealing to the extended D loop leads to the formation of Holliday junctions, which is mediated by RAD52 [40, 51].

R37 Resolution of D loop by RTEL1 producing non-crossover product

Regulator of Telomere Length 1 (RTEL1) is an essential helicase that is essential in DNA repair as well as telomere maintenance. It promotes the disassembly of D loop at an early stage of HR generating non-crossover product, and thus promotes synthesis-dependent strand annealing (SDSA), one of the HR-mediated repair mechanisms [38, 48].

R38 Formation of MUS81-EME1 Complex

See description for R39.

R39 Resolution of D loop by MUS81-EME1 Complex producing crossover product

MUS81 is a member of the XPF family of heterodimeric nuclease and can form a complex with EME1. The major role of this complex is to resolve Holliday Junction, but it also can efficiently cleaves D loop, leading to crossover product [54, 55].

R40 Formation of BLM, TOPOIIIα, RMI1 and RMI2 Complex

See description for R41.

R41 Dissolution of Holliday junctions by BLM, TOPOIIIα, RMI1 and RMI2 Complex giving non-crossover product

BLM, the helicase mutated in a cancer predisposition syndrome known as Bloom Syndrome, forms a complex with topoisomerase 3α, RMI1 (also known as BLAP75) and RMI2 (also known as BLAP18). This complex stimulates the dissolution of Holliday structure and always generate non-crossover product. It is therefore important for preventing aberrant recombination, elevated sister chromatid and genome rearrangements [38, 56].

R42 Resolution of Holliday junctions by GEN1 giving noncrossover product

See description for R43.

R43 Resolution of Holliday junctions by GEN1 giving crossover product

GEN1 is a member of the Rad2/XPG family of monomeric, structure-specific nucleases, but it has been adapted from a simple 5'-flap endonuclease into an HJ resolvase, generating crossover or non-crossover product [38, 57].

R44 Resolution of Holliday junctions by MUS81-EME1 complex giving crossover product

See description for R39.

R45 Formation of SLX1-SLX4 Complex

See description for R47.

R46 Resolution of Holliday junctions by SLX1-SLX4 C giving noncrossover product

See description for R47.

R47 Resolution of Holliday junctions by SLX1-SLX4 C giving crossover product

SLX4 is an important Fanconi Anemia protein also known as FANCP. It associates with SLX1 to form a complex that has remarkable Holliday junction-resolving activity, generating crossover or non-crossover product [38, 58].

R48 RAD52 binds ssDNA to initiate SSA

RAD52-dependent single-strand annealing (SSA) subpathway of HR is an error-prone but efficient way to repair DSBs between two direct repeat sequences, each present on one side of the DSB ends. SSA is an important pathway for both DNA repair and mutagenesis given the fact that almost half of the human genome consists of repeated sequences. This pathway is initiated by RAD52 that binds 3' ssDNA ends and functions in concert with RPA to facilitate strand annealing between the two direct repeats [38, 59, 60].

R64 Heterodimer formed by ERCC1 and XPF

See description for R49.

R49 Removal of non-homologous 3' single-stranded flaps at DSB ends by ERCC1 and XPF

XPF and ERCC1 form a heterodimer that functions as a 5′-3′ structure-specific endonuclease. The key activity of this heterodimer is to remove non-homologous 3′ single-stranded flaps at broken ends before they are rejoined (in this case, it is the sequence between the two repeats). In addition to SSA, this heterodimer plays an important role in another a few DNA repair pathways, including Nucleotide Excision Repair (NER), Fanconi Anemia (FA) pathway, and Microhomology-Mediated End-Joining pathway (MMEJ, a subpathway of Non-homologous End-joining repair pathway) [61, 62].

R50 Ligation of the DSB ends by LigIII

The last step of SSA is the ligation of two DSB ends to restore the broken DNA as a continuous duplex. This process is conducted by DNA ligase III. [63].

R51 SUMOylation of 53BP1 by PIAS4

The tumor suppressor 53BP1 is SUMOylated by SUMO ligases PIAS4 in response to DSB, which is essential for the association of 53BP1 with the damage site [64, 65].

R52 SUMOylation of BRCA1 by PIAS1

In response to DSB, tumor suppressor BRCA1 is SUMOylated by PIAS1 or PIAS4, two SUMO ligases that are important for DDR. This reaction stimulates the ubiquitin ligase activity of BRCA1 and promotes the recruitment of BRCA-A complex to the damage site [38, 64, 65]

R58 SUMOylation of BRCA1 by PIAS4

See description for R52.

R53 BRCA2 stabilized by RBMX

RBMX is a heterogeneous nuclear ribonucleoprotein that has a role in alternative splicing. It has been recently shown that RBMX is recruited to DNA damage site with the help of PARP1 and promotes HR by facilitating proper expression of BRCA2 [66].

R54 Phosphorylation of MDC1 by CK2

It has been shown that MDC1 is phosphorylated by casein kinase 2 (CK2) on a cluster of conserved repeat motifs. Mutation of these conserved motifs in MDC1 or depletion of CK2 disrupts the interaction between MDC1 and NBS1 (one of the MRN complex components), and in turn abrogates the targeting of MRN on DNA DSB sites [10, 11, 38].

R55 Association of phosphorylated MDC1 with MRN complex

See description for R54.

R56 Formation of BRCA2-DSS1 complex

DSS1 forms a complex with BRCA2, which is required for the stability and ssDNA binding ability of the latter protein. DSS1 also facilitates BRCA2 in RAD51–ssDNA filament formation [21, 67].

R59 SUMOylation of RNF168 by PIAS4

It has been recently shown that RNF168 is SUMOylated by PIAS4 in response to DSB, and this modification might be important for RNF168 retention at the DSB site [68, 69].

R60 SUMOylation of HERC2 by PIAS4

The ubiquitin ligase HERC2 is SUMOylated by PIASA in response to DSB. This modification of HERC2 promotes its interaction with RNF8 and is necessary for stabilizing the RNF8-UBC13 complex [69].

R61 Heterodimer formed by RNF20 and RNF40

See description for R63.

R62 Phosphorylation of RNF20 and RNF40 by ATM

See description for R63.

R63 Monoubiquitylation of histone H2B by RNF20 and RNF40 dimer

RNF20 and RNF40 are both E3 ubiquitin ligase and form a heterodimer termed BRE1. This complex plays a role in HR by monoubiquitinating histone H2B. This process relies on phosphorylation of RNF20 and RNF40 and is essential for DNA end resection in HR and recruitment of downstream HR proteins such as RAD51 and BRCA1 [70-72].

R64 SUMOylation of MDC1 by PIAS4

See description for R65.

R65 Ubiquitination of SUMOylated MDC1 by RNF4

It has been recently shown that MDC1 is SUMOylated mainly by PIAS4 at sites of DNA damage, which is then recognized and ubiquitinated by the SUMO-targeted E3 ubiquitin ligase RNF4. This process is required for MDC1 degradation and removal of MDC1 and 53BP1 from the damage sites, and is important for the recruitment of downstream HR proteins such as CtIP, RAD51 and RPA [73-75].

A1.2 Non-homologous end-joining

Non-homologous end-joining (NHEJ) repair is the main pathway for repairing DNA double-strand breaks (DSBs). It functions in all phases of the cell cycle but predominates in G0 and G1 phase. In contrast to Homologous Recombination (HR) repair, NHEJ directly rejoins the DSBs without the need for homologous template - namely, it restores genomic integrity without ensuring sequence fidelity - and thus be considered as error-prone [76-78]. NEHJ has two subpathways – canonical NEHJ (C-NHEJ) and alternative NHEJ (A-NHEJ). The choice between them is regulated by both 53BP1 (which promotes C-NHEJ) and PARP1 (which promotes A-NHEJ), and A-NHEJ usually operates under C-NHEJ defective conditions [79, 80]. Interestingly, it was recently shown that deregulated NHEJ plays a critical role in the hypersensitivity of HR-deficient cells to PARP inhibitors [81].

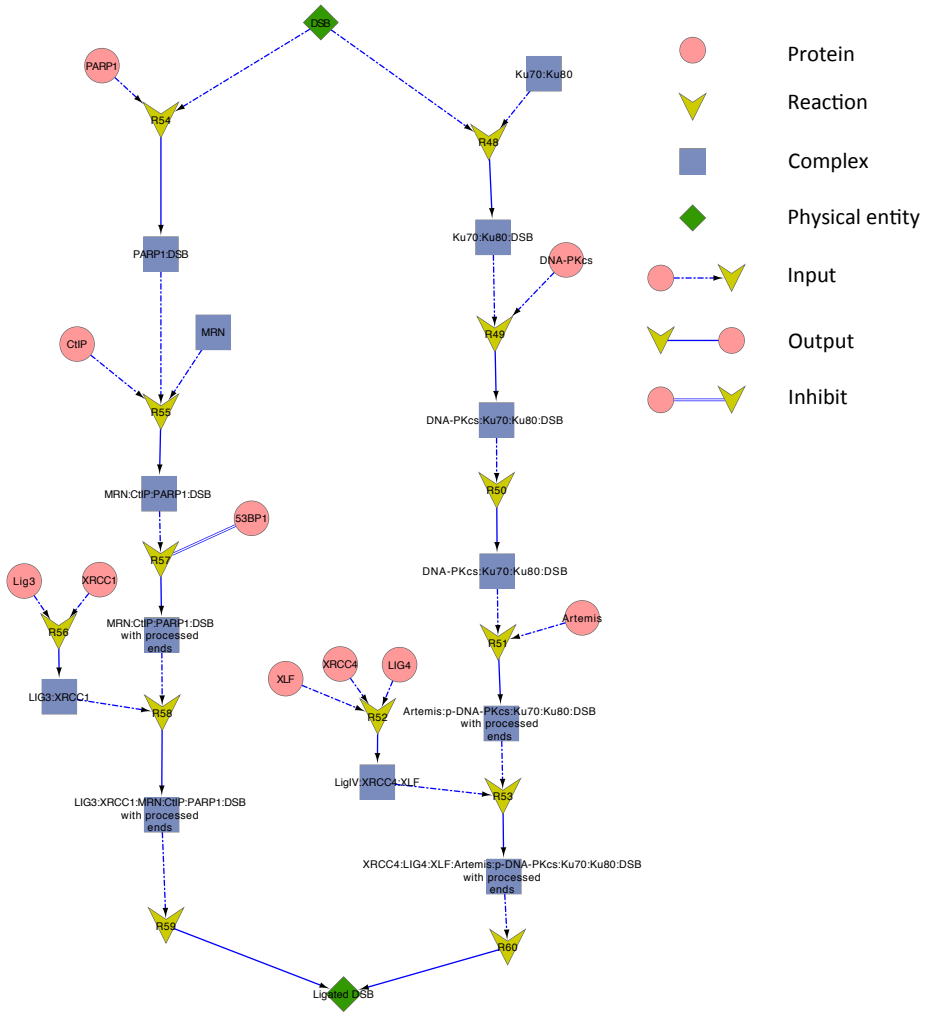


Figure 2 Non-homologous end joining (NHEJ) pathway figure. The original PDF version is deposited at UQ eSpace.

A1.2.1 Reaction Description

R48 Binding of Ku70-Ku80 heterodimer with ends of DNA double-strand break (DSB)

Ku is a heterodimer consisting of Ku70 and Ku80, which functions as the major DSB sensing protein in NHEJ, and also possesses a DNA end processing activity. In response to DSB, it rapidly binds to each of the two broken ends with high affinity but without apparent sequence specificity, and forms a Ku-DNA complex that serves as a docking site for the subsequent NHEJ proteins [76, 82, 83].

R49 Recruitment of DNA-PKcs to DNA damage sites

Binding of Ku to DSB ends is followed by recruitment of the DNA-dependent protein kinase catalytic subunit (DNA-PKcs) that forms a complex called DNA-PK with Ku. Formation of this active kinase complex causes the phosphorylation of other NHEJ proteins as well as the autophosphorylation of DNA-PKcs. This autophosphorylation reaction is thought to induce a conformational change in DNA-PKcs, resulting in the release of the autophosporylated DNA-PKcs from DNA before or after end processing [76, 83, 84].

R50 Autophosphorylation of DNA-PKcs

See description for R49.

R51 Recruitment of Artemis to damage sites where it performs DSB ends processing

DSB ends often need to be processed before ligation, and this is performed by the 5'-3' endonuclease Artemis in NHEJ. It has been proposed that the recruitment of Artemis to damage site and the activation of its nucleolytic activity requires the autophosphorylation of DNA-PKcs [76, 83, 85].

R61 Gap filling by DNA polymerase μ

DNA polymerase μ (Pol μ) and DNA polymerase λ (Pol λ) are polymerases of X family, which are recruited to DSBs via their interactions with Ku and responsible for fill-in synthesis during NHEJ. These two polymerases are both prone to slippage on the template strand and as a result, repeats are commonly seen at NHEJ junctions. In addition, Pol μ may has a ability for

template-independent synthesis; that is, it can cross from one DNA end to another DNA end [76, 83, 86].

R62 Gap filling by DNA polymerase λ

See description for R61.

R52 Formation of LIGIV:XRCC4:XLF complex

See description for R60.

R53 Recruitment of LIGIV:XRCC4:XLF complex to damage sites

See description for R60.

R60 Ligation of DSB ends by LIGIV:XRCC4:XLF complex

The last step in C-NHEJ is the ligation of DSB ends once they have been processed, which is catalyzed by a complex composed of DNA ligase IV (LIGIV) and another two nonenzymatic components named XRCC4 and XLF. In this complex, XRCC4 seems to stabilize LIGIV and stimulates its joining activity [87, 88]; XLF may help LIGIV to ligate a more diverse array of DNA [89]. This complex is targeted to damage sites by an interaction with Ku [90, 91].

R54 Binding of PARP1 with DSB ends

A-NHEJ is another form of NHEJ that involves more resection of the free DNA ends to find microhomologies, and thus is less accurate than C-NHEJ. This pathway is initiated by PARP1 poly (ADP-ribose) polymerase 1 (PARP1), which competes with Ku for DSB end binding. As Ku has much higher affinity for DNA ends than PARP1, the activities of A-NHEJ is limited to cells that are deficient in C-NHEJ [80, 92, 93].

R55 Recruitment of MRN complex and CtIP to the DSB site

See description for R57.

R57 DSB ends processed by MRN and CtIP, which could be inhibited by 53BP1

Following binding of PARP1 to DSB ends, MRN and CtIP are recruited to DSB sites and conduct single-stranded end-resection during A-NHEJ, which exposes microhomologies that

promote pairing of broken ends [94, 95]. This process could be blocked by 53BP1, which promotes C-NHEJ to increase NHEJ accuracy [79].

R56 Formation of LIG3:XRCC1 complex

See description for R59.

R58 Recruitment of the LIG3:XRCC1 complex to the DSB site

See description for R59.

R59 DSB ends ligated by the LIG3:XRCC1 complex

The ligation step of A-NHEJ is performed by a complex consisting of DNA ligase III (LIGIII) and XRCC1. The ligation function of this complex is regulated by PARP1. Interestingly, LIG3, XRCC1 and PARP1 are also involved in Base Excision Repair (BER) [96-98].

A1.3 Nucleotide excision repair

Nucleotide excision repair (NER) deals with helix-distorting damage, such as lesions induced by UV light. It is often subclassified into transcription-coupled NER (TCR), which recognizes and removes helical distortions selectively from the transcribed strand of active genes; and global-genome NER (GGR), which is able to repair lesions throughout the entire genome. Impairment in this DNA repair activity has been associated with several human diseases, including xeroderma pigmentosum (XP), a hereditary syndrome characterized by UV hypersensitivity and skin cancer [99, 100].

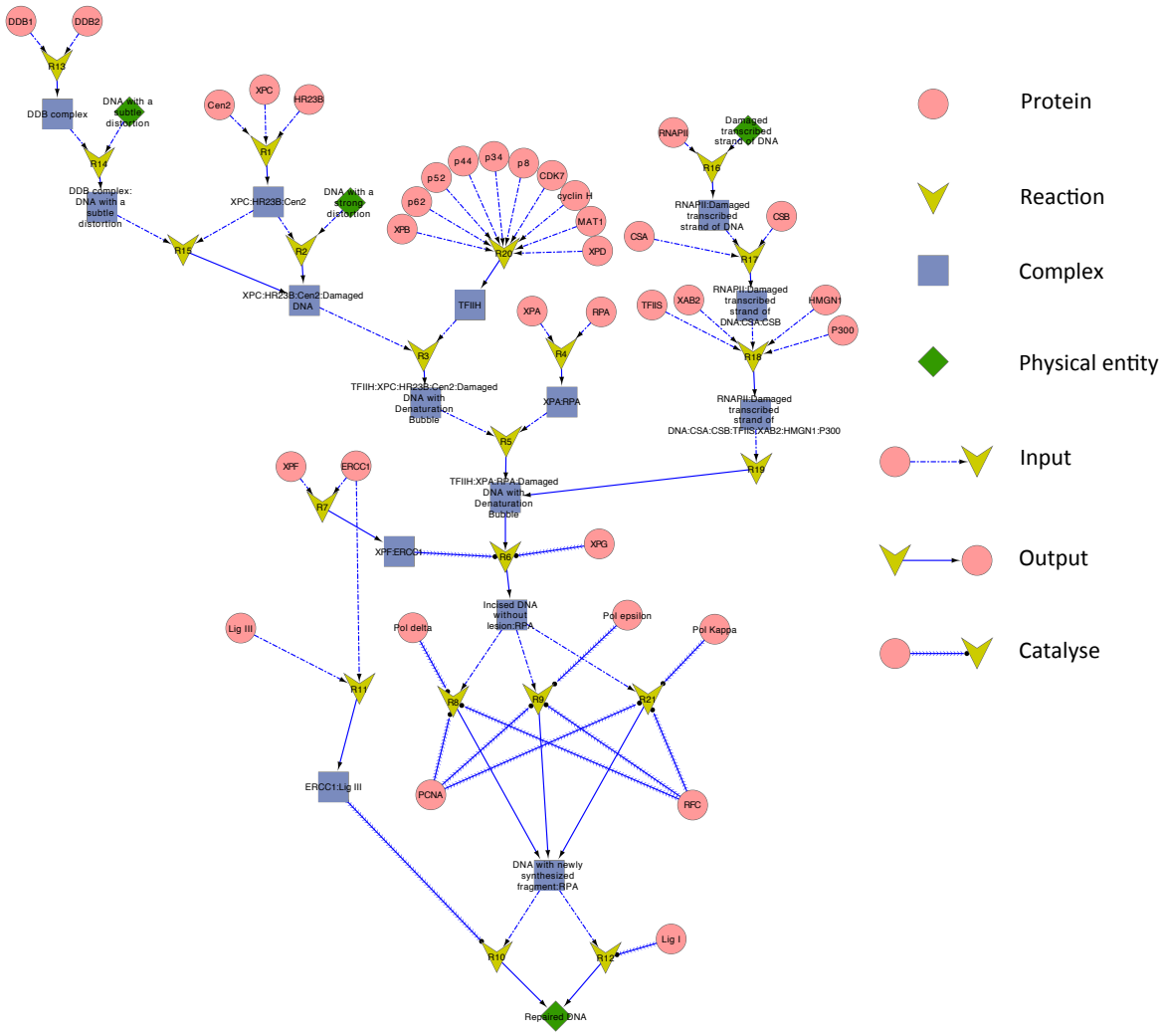


Figure 3 Nucleotide excision repair (NER) pathway figure. The original PDF version is deposited at UQ eSpace.

A1.3.1 Reaction Description

R1 Formation of XPC-HR23B-Cen2 complex

See description for R2.

R2 Recognition of DNA with a strong distortion by the XPC-HR23B-Cen2 complex

In GGR, the majority of the DNA lesions are detected by the XPC-HR23B-Cen2 complex which consists of three subunits: XPC, HR23B and Cen2. XPC is a DNA binding protein with a strong preference for damaged DNA. It plays the distortion-sensing role in the complex and is also essential for the recruitment of downstream NER factors. HR23B is one of the two human orthologs of the yeast protein Rad23 (the other one is HR23A). Although these two proteins are both able to interact with XPC and increase its activity in NER, it is HR23B that is commonly found in

association with XPC. The third component of the complex, Cen2, serves to stabilize the complex and improve its activity in NER [100-102].

R20 Formation of TFIIH complex

TFIIH has an important role in both transcription initiation and NER. It is a multiprotein complex with ten subunits, 6 of which (XPB/ERCC3, p62, p52, p44, p34 and p8) form the core complex and 3 of which (CDK7, cyclin H and MAT1) form the CAK complex. The remaining subunit, XPD (also named ERCC2), connects these two subcomplexes by interacting with the p44 and MAT1 of the core or the CAK subcomplex, respectively. The precise function of all these subunits is not known, but it is thought that CDK7 acts as a cyclin-dependent kinase and XPB and XPD are ATP-dependent helicases of opposite polarities [100, 103].

R3 A denaturation bubble of about 30 nucleotides around the lesion opened by TFIIH

Following the XPC-HR23B-Cen2 complex binding to the DNA lesions, TFIIH is recruited through the interaction of XPC with at least two TFIIH subunit (p62 and XPB). It then mediates the excision of the damaged DNA by unwinding the DNA to open a denaturation bubble around the lesion. XPB and XPD, the two subunits of TFIIH also help to stabilize the single-stranded DNA during the repair process [100, 102, 103].

R4 Formation of XPA-RPA complex

See description for R5.

R5 Recruitment of RPA and XPA to DNA lesions helps to release the XPC-HR23B-Cen2 complex and confirm the presence of DNA damage.

The XPA protein forms a complex with RPA complex, which plays an indispensable role in NER. In particular, RPA stabilizes the unwound state of DNA by binding to the undamaged DNA strand; XPA binds specifically to damaged DNA, which is reinforced by interaction with RPA, and has been implicated in the damage-verification step of NER [102, 104, 105].

R6 Dual incision of damaged DNA by XPF-ERCC1 complex and XPG

ERCC1- XPF complex and XPG are both structure-specific endonucleases. The recruitment of ERCC1–XPF to DNA lesion sites is thought to be mediated by both ERCC1/XPA and XPF/RPA interactions, whereas the recruitment of XPG depends on TFIIH. After being recruited, ERCC1-XPF and XPG incise the damaged DNA strand 5′ and 3′, respectively, to the lesions [61, 100, 106].

R7 Formation of XPF-ERCC1 complex

XPF and ERCC1 form a complex that functions as a structure-specific endonuclease. This heterodimer has important roles in a few DNA repair pathways, including NER, Fanconi Anemia pathway, Single-strand Annealing pathway (SSA, a subpathway of Homologous Recombination pathway) and Microhomology-mediated End-joining pathway (MMEJ, a subpathway of Non-homologous End-joining pathway) [61, 62].

R8 Gap filling by DNA polymerase delta facilitated by PCNA and RFC

Following the removal of the damage-containing oligonucleotide, the single-stranded gap left behind is filled by the replicative DNA polymerase delta (Pol δ), DNA polymerase epsilon (Pol ε) or the translesion DNA polymerase Kappa (Pol κ). This step depends on PCNA, which stimulates and coordinates the polymerase activities, and RFC, which helps loading of PCNA onto the DNA template. The recruitment of RFC and PCNA seems to be dependent on RPA and XPG [100, 102, 105].

R9 Gap filling by DNA polymerase epsilon facilitated by PCNA and RFC

See description for R8.

R10 Gap sealing mainly by the DNA ligase III-ERCC1 complex

See description for R11.

R11 Formation of ERCC1-DNA ligase III complex

DNA ligase III forms a complex with ERCC1, which is required for the normal function of DNA ligase III [107].

R12 Gap sealing by DNA ligase I in replication cells

Following DNA synthesis, the remaining nick is rejoined by DNA ligase I or DNA ligase III-XRCC1 complex to conclude NER [100, 108, 109].

R13 Formation of DDB complex

See description for R15.

R14 DNA with a subtle distortion recognized by DDB complex

See description for R15.

R15 Recruitment of XPC-HR23B-Cen2 complex with the help of DDB complex

DDB complex is a damage-sensor in GGR, which monitor DNA lesions that only cause subtle helix distortion and thus are difficult to be recognized by the XPC-HR23B-Cen2 complex. DDB is a heterodimer consisting of DDB1 and DDB2, and may have evolved especially to cope with dinucleotide lesions, such as UV-induced photodimers. It also seems to promote the recruitment of XPC at sites of DNA damage [100, 102, 105].

R16 Initiation of TCR by RNAPII

TCR is initiated by RNA polymerase II (RNAII) when this polymerase encounters a DNA lesion in the transcribed strand during transcription and so gets stalled. Damage-arrested RNAII also serves as a trigger to assemble other TCR proteins [110-112].

R17 Recruitment of protein CSB and CSA

The damage-stalled RNAPII leads to the recruitment of protein CSB which tightly binds to RNAII and function as a repair coupling factor to attract the remaining NER factors. Protein CSA is also recruited in a CSB-dependent way, and in cooperation with CSB, CSA is required to recruit other NER factors (such as XAB2, HMGN1 and TFIIS) to damage site [100, 111, 112].

R18 Recruitment of additional TCR-specific factors: TFIIS, XAB2, HMGN1 and P300

Following the recruitment of CSA and CSB to the damage site, NER proteins XAB2, HMGN1 and TFIIS are recruited in a CSA-dependent way while protein p300 is recruited in a CSB-dependent way. It has been suggested that the recruitment of p300 and HMGN1 might facilitate chromatin remodeling and reverse translocation of RNAPII, therefore allowing the removal of the blocking damage by the repair machinery and the resumption of transcription [100, 111, 112].

R19 Formation of TFIIH-XPA-RPA-Damaged DNA with denaturation bubble complex

Following damage detection, GGR and TCR merge into a common mechanism to unwind the DNA around the lesion, incise the DNA at both sides of the lesion and fill the resulting gap [100, 111, 112].

A1.4 Base excision repair

Base Excision Repair (BER) is primarily responsible for removing small, non-helix-distorting base lesions caused by oxidation, alkylation and deamination. It uses a collection of specific enzymes (known as glycosylase), to recognize and remove different types of damaged or inappropriate bases, forming apurinic/apyrimidinic (AP) sites [113, 114].

The AP sites are further processed by AP-endonuclease 1 (APE1) that generates DNA Single Strand Breaks (SSBs) with a 5'-sugar phosphate. This 5’-sugar phosphate is then cleaved by DNA polymerase β (Pol β), which at the same time adds one new nucleotide to the 3' end of the nick. The nick is finally sealed by DNA ligase III. This process is termed short-patch BER [113, 114].

If the 5'-sugar phosphate is resistant to cleavage by Pol β, then after the addition of the first new nucleotide by Pol β, DNA polymerase delta or epsilon comes in and adds 2-8 more nucleotides into the repair gap, resulting a flap structure that is then removed by the Flap Endonuclease-1 (FEN-1) in a PCNA-dependent manner. In the final step, the remaining nick is sealed by DNA ligase I. This process is termed long-patch BER [113, 114].

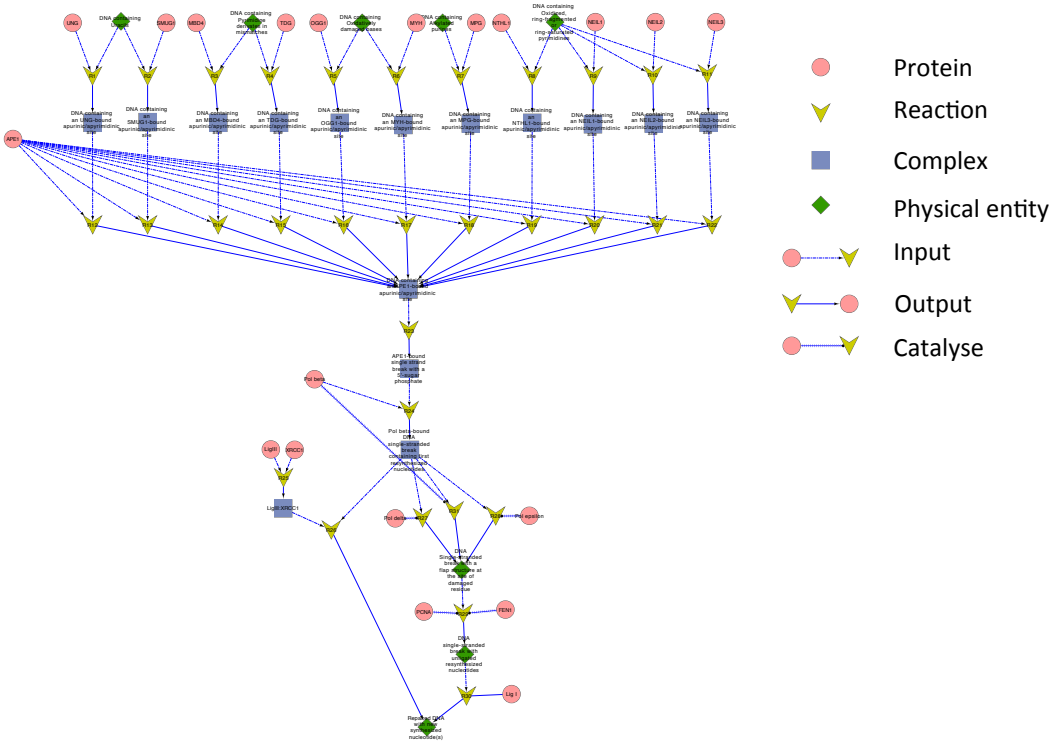


Figure 4 Base excision repair (BER) pathway figure. The original PDF version is deposited at UQ eSpace.

A1.4.1 Reaction Description

R1 – R11 description

DNA glycosylase refers to a large family of enzymes that cleave the damaged base from the sugar-phosphate backbone. They are classified as monofunctional and bifunctional, depending on their reaction mechanisms. The monofunctional glycosylases (e.g. UNG) only have glycosylase activity that cleaves the glycosidic bond linking the damaged base and the ribose. The bifunctional glycosylases (e.g. OGG1) have both glycosylase and AP lyase activity. In addition to cleaving the glycosidic bond, they also be able to incise the DNA backbone 3' to the AP site [113-115].

R1 Binding and cleavage of Uracils by monofunctional glycosylase UNG

R2 Binding and cleavage of Uracils by monofunctional glycosylase SMUG1

R3 Binding and cleavage of Pyrimidine derivates in mismatches by monofunctional glycosylase MBD4

R4 Binding and cleavage of Pyrimidine derivates in mismatches by monofunctional glycosylase TDG

R5 Binding and cleavage of Oxidatively damaged bases by bifunctional glycosylase OGG1

R6 Binding and cleavage of Oxidatively damaged bases by monofunctional glycosylase MYH

R7 Binding and cleavage of Alkylated purines by monofunctional glycosylase MPG

R8 Binding and cleavage of Oxidized, ring-fragmented or ring-saturated pyrimidines by bifunctional glycosylase NTHL1

R9 Binding and cleavage of Oxidized, ring-fragmented or ring-saturated pyrimidines by bifunctional glycosylase NEIL1

R10 Binding and cleavage of Oxidized, ring-fragmented or ring-saturated pyrimidines by bifunctional glycosylase NEIL2

R11 Binding and cleavage of Oxidized, ring-fragmented or ring-saturated pyrimidines by bifunctional glycosylase NEIL3

R12 – R23 Reaction Description

Following the generation of the AP site, the DNA glycosylase is displaced by APE1 that binds to the AP site and hydrolyzes the phospho-diester bond 5' to the this site, generating a DNA single strand breaks (SSBs) with a 5'-terminal sugar phosphate [113, 114, 116].

R12 Displacement of glycosylase by APE1

R13 Displacement of glycosylase by APE1

R14 Displacement of glycosylase by APE1

R15 Displacement of glycosylase by APE1

R16 Displacement of glycosylase by APE1

R17 Displacement of glycosylase by APE1

R18 Displacement of glycosylase by APE1

R19 Displacement of glycosylase by APE1

R20 Displacement of glycosylase by APE1

R21 Displacement of glycosylase by APE1

R22 Displacement of glycosylase by APE1

R23 APE-mediated DNA backbone incision 5' to the AP site

R24 Displacement of APE1 and adding one new nucleotide by Pol β

Following the APE1-mediated cleavage, it is then displaced from the damaged DNA by DNA polymerase III, which also incorporates one nucleotide to the 3’-end of the arising single-nucleotide gap [113, 117, 118].

R25 Formation of LigIII-XRCC1 complex

See description for R26.

R26 DNA nick ligation by DNA ligase III-XRCC1 complex

If the 5'-terminal sugar phosphate generated by APE1 cleavage is not resistant to the β-elimination reaction mediated by Pol β, it can be removed from the single-stranded break, and the final nick can be sealed by DNA Ligase III with the help of the scaffold protein XRCC1 [114, 119, 120]. This process is commonly referred to as short-patch BER [113, 114, 118].

R27 Strand-displacement DNA synthesis by Pol Delta

R28 Strand-displacement DNA synthesis by Pol Epsilon

R31 Strand-displacement DNA synthesis by Pol beta

If the 5'-terminal sugar phosphate generated by APE1 cleavage is resistant to the β-elimination, Pol β may or may not dissociate from the damage site after adding the first nucleotide into the repair gap. Then, either Pol β, Pol δ or Pol ε will add 2-8 more nucleotides into the repair gap, a process known as "strand-displacement DNA synthesis", producing a multi-nucleotide repair patch and a single-stranded DNA structure termed DNA flap. [117, 118, 121].

R29 Removal of DNA flap structure by FEN1 and PCNA

In this step, the DNA flap structure generated by DNA polymerase, which contains the AP site and a few displaced unannealed nucleotides, is cleaved by FEN1 in a PCNA-dependent manner [113, 118, 122].

R30 DNA nick ligation by DNA ligase I

Following the flap cleavage by FEN1 and PCNA, the remaining nick in the DNA backbone is sealed by DNA ligase I. This process is commonly referred to as long patch BER [113, 114, 118].

A1.5 Mismatch repair

The Mismatch Repair (MMR) pathway is the main mechanism responsible for the repair of base-base mismatches and insertion/deletion loops (IDL) that are formed during DNA replication. It degrades the error-containing section of the newly synthesized strand and therefore provides the DNA polymerase with another chance to generate an error-free copy of the template sequence. In

MMR-deficient tumor cells, mutation rates are 100 to 1,000 fold greater in comparison to normal cells [123-125].

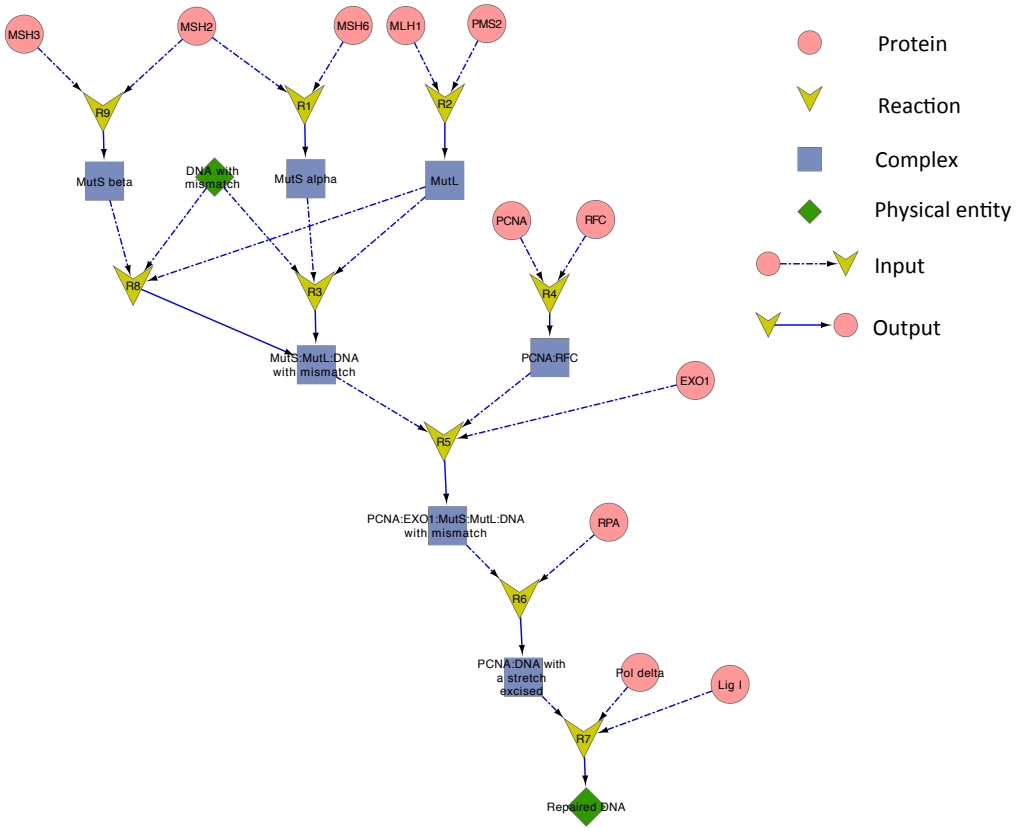


Figure 5 Mismatch repair (MMR) pathway figure. The original PDF version is deposited at UQ eSpace.

A1.5.1 Reaction Description

R1 Formation of MutSα complex

Human MutS heterodimer is present in two basic forms (MutSα and MutSβ), and initiates the MMR pathway. MutSα encompasses MSH2 in a complex with MSH6, and primarily responsible for the repair of base substitutions and small mismatched loops. MutSβ consists of MSH2 and MSH3, and repairs both small loops as well as large loop mismatches (~10 nucleotide loops) [125-127].

R9 Formation of MutSβ complex

Described in R1

R2 Formation of MutLα complex

Human MutL heterodimer is present in three basic forms - MutLα, MutLβ and MutLγ. MutLα is made up of MLH1 and PMS2, which is responsible for the primary activity of MutL during MMR. MutLβ, which consists of MLH1 and PMS1, might contribute a minor role. The function of MutLγ (MLH1 and MLH3) remains unknown [124, 125, 127].

R3 DNA mismatch detection and recruitment of MutL by MutSα

Described in R8

R8 DNA mismatch detection and recruitment of MutL by MutSβ

MutS (either MutSα or MutSβ) initiates the MMR pathway by recognizing distortions in the DNA double helix structure caused by mismatched bases. After binding to double-strand DNA at damage site, it then recruits MutL that may function as a mediator for the interactions of downstream MMR proteins. MutS and MutL also form a complex that leaves the mismatch site and slides up and down the flanking DNA sequence until it encounters a single-strand DNA gap bound by PCNA and RFC [124, 125, 127].

R4 Formation of PCNA-RFC complex

Protein PCNA is a cofactor required for DNA synthesis mediated by DNA polymerase δ. It has an important role in a few DNA repair pathways, including MMR, Base Excision Repair (BER) and Nucleotide Excision Repair (NER). In MMR, it forms a complex with protein RFC that facilitates the loading of PCNA onto primed DNA templates [125, 127, 128].

R5 Displacement of RFC and recruitment of EXO1

As MMR is mainly responsible for replication error repair, it must be able to distinguish parental DNA from daughter DNA. It has been suggested that when the sliding clamp of MutS and MutL encounters PCNA and RFC, RFC is displaced, which allows exonuclease EXO1 to access the daughter strand DNA [125, 127, 129].

R6 Degradation of DNA across the site of mismatch by EXO1 and with the help of RPA

With the guidance of MutS and MutL complex, EXO1 removes the sequences across the mismatch site on daughter strand, and is then inactivated by MutL. While the daughter strand is processed, the parental strand of DNA is stabilized by RPA [125, 127, 129].

R7 New DNA synthesis and gap ligation by DNA polymerase delta and DNA ligase I

Once the mismatched DNA on daughter is excised, DNA polymerase delta synthesizes new DNA followed by DNA ligase I that seals the remaining nick to complete the repair process [125, 127, 129].

A1.6 Fanconi anemia

Fanconi Anemia (FA) is a rare hereditary genomic instability syndrome characterized by bone marrow failure, developmental abnormalities, and cancer predisposition. This disorder is caused by mutations in genes that are responsible for the removal of DNA Interstand Crosslinks (ICLs). To date, 15 FA or FA-like genes (FANCA, B, C, D1, D2, E, F, G, I, J, L M, N, P and O) and two other FA-associated genes, FAAP24 and FAAP100, have been identified. The pathway that they constitute is therefore termed the Fanconi Anemia (FA) pathway. The FA pathway coordinates three common DNA repair pathways, including Homologous Recombination (HR), Nucleotide Excision Repair (NER) and Translesion DNA synthesis (TLS). In addition, most of the FA and associated proteins form a large ubiquitin ligase complex termed the FA core complex [130, 131].

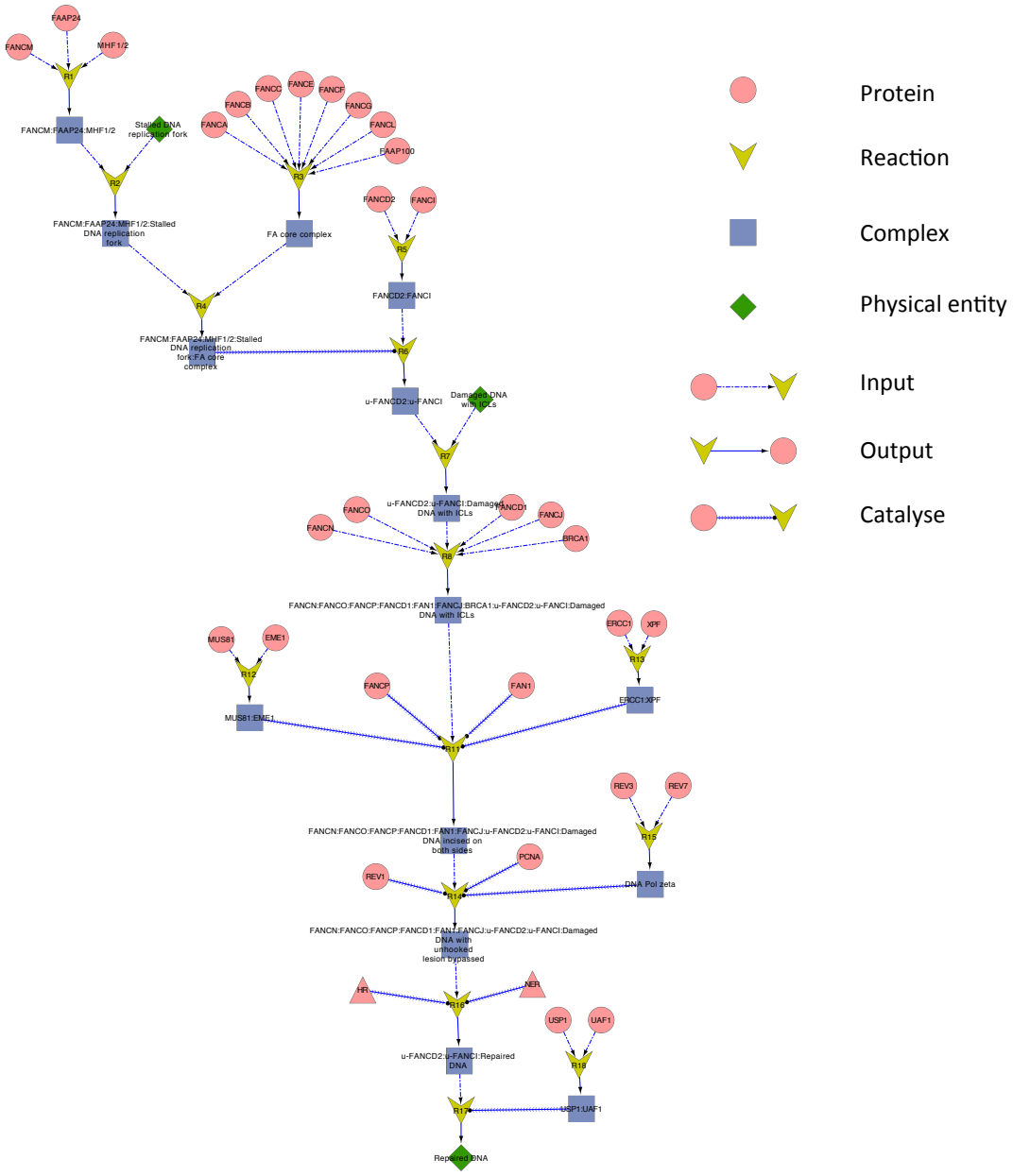


Figure 6 Fanconi Anaemia (FA) pathway figure. The original PDF version is deposited at UQ eSpace.

A1.6.1 Reaction Description

R1 Formation of FANCM-FAAP24-MHF1/2 complex

The FANCM protein forms a complex with FAAP24, MHF1 and MHF2, and initiates the FA pathway. FAAP24 helps FANCM to recognize DNA lesion, recruit the FA core complex, stabilize the stalled replication fork, and initiate ATR mediated checkpoint signaling. MHF1 and MHF2 maintain the stable association of FANCM with chromatin and promote efficient pathway activation [131-133].

R2 Recognition of stalled replication fork by FANCM-FAAP24-MHF1/2 complex

When a replication fork encounters an ICL during replication, the replication fork arrests near the lesion, resulting in aberrant DNA structures. The FANCM– FAAP24–MHF1/2 complex recognizes the stalled replication fork structure and recruits the FA core complex to the ICL region. FANCM can also prevent the collapse of replication fork via its translocase activity [131-133].

R3 Formation of FA core complex

Eight FA proteins (FANCA/B/C/E/F/G/L/M) form a multisubunit nuclear complex, known as the FA core complex, which through FANCL, acts as an ubiquitin E3 ligase to mono-ubiquitinate FANCD2 and FANCI following DNA damage [131, 134, 135].

R4 FA core complex recruited to the DNA lesion by direct interaction between FANCM and FANCF

FANCM is an essential component of the FA core complex. When an ICL is present, it is recognized by FANCM– FAAP24–MHF1/2 complex, which recruits the rest of the FA core proteins by interaction between FANCM and FANCF [131, 136].

R5 Formation of FANCD2-FANCI complex

FANCD2 and FANCI form a complex (also known as ID complex) that can be monoubiquitinated by the FA core complex before being localized to chromatin in response to DNA damage. Moreover, the ubiquitination of each of these two proteins is important for the maintenance of ubiquitin on the other [131, 137, 138].

R6 Monoubiquitination of FANCD2 and FANCI on chromatin by FA core complex

See the description for R7.

R7 Recruitment of monoubiquitinated FANCD2 and FANCI to the damaged DNA with ICLs

The Monoubiquitination of FANCD2 and FANCI by the FA core complex is the key regulatory step in the FA pathway. Following this mono-ubiquitination, the ubiquitin-tagged FANCD2–FANCI complex is relocalized to the DNA lesion where it coordinates cross-link repair

activities together with downstream FA proteins. In particular FANCD2 acts as a landing pad to recruit multiple nucleases such as FAN1 and FANCP (SLX4) in order to initiate the nucleolytic incision flanking the ICL [131, 134, 139].

1.1. R8 Recruitment of other DNA repair factors to the damaged site

Following the recruitment of FANCD2-FANCI complex to DNA lesions, multiple downstream FA proteins, including FANCD1 (also known as BRCA2), FANCJ (also known as BRIP1), FANCN (also known as PLAB2), FANCO (RAD51C), FANCP (also known as SLX4), BRCA1 and FAN1, are also recruited. FAN1 and FANCP act as a DNA nuclease in ICL repair in order to initiate nucleolytic incision; BRCA1, FANCD1, FANCJ, FANCN, FANCO play an important role in the HR process that is involved in ICL repair [130, 131, 135, 138, 140];

1.2. R12 Formation of MUS81-EME1 complex

MUS81 and EME1 form a complex that acts as a endonuclease and plays an important role in rescuing stalled replication forks during ICL repair, and resolving the Holliday Junctions (HJs) in eukaryotes [54, 131, 134].

1.3. R13 Formation of ERCC1-XPF complex

The complex formed by ERCC1 and XPF is an endonuclease that is essential for NER and has important roles in ICL repair and DSB repair. Therefore, it has a critical role in the response of cancers to a range of DNA-damaging chemotherapeutics [61, 62, 131].

R11 Dual incision on each side of the ICL performed by MUS81-EME1 and ERCC1-XPF respectively

During ICL repair, the nucleolytic dual incisions on each side of the ICL (also known as unhooking) are performed by MUS81-EME1 and ERCC1-XPF, respectively. This unhooking process converts a stalled replication fork into a DSB. FANCP and FAN1 may also contribute to this process [131, 134, 141].

R15 Formation of the DNA Pol zeta complex

DNA polymerase ζ is a complex consisting of REV 3 and REV7. It belongs to a family known as translesion synthesis polymerases which are low fidelity DNA polymerases that allow cells to replicate over the replication-blocking lesions without correcting it [131, 134, 142].

R14 The unhooked lesion bypassed by DNA Pol ζ and REV1

Following ICL unhooking, the cross-linked nucleotides are bypassed by TLS polymerase REV1 and Pol ζ to restore a nascent strand that is subsequently used as a template of HR. PCNA also contributes to this process by recruiting and regulating REV1 and DNA polymerase ζ to the damage site [131, 138, 143].

R16 The participation of HR and NER in ICL repair

The dual incision flanking the ICL region creates a DSB as an intermediate in the ICL repair process, which is then repaired by HR using the homologous template that has been repaired by TLS. The downstream FA proteins promote RAD51-dependent strand invasion and the resolution of recombinant intermediates. The remaining DNA adducts are removed and gap filled by NER [131, 138, 144].

R18 Formation of USP1-UAF1 complex

See the description for R17.

R17 Deubiquitination and release of FANCD2 and FANI by the USP1-UAF1 complex

The deubiquitinating enzyme USP1 and its activating partner, UAF1, forms a complex that removes the monoubiquitin from FANCD2 - FANCI complex and completes the ICL repair. In animal models, knockdown either USP1 or UAF1 can lead to hypersensitivity to DNA cross-linking agents [131, 145, 146].

References

1. Lamarche BJ, Orazio NI, Weitzman MD: **The MRN complex in double-strand break repair and telomere maintenance.** *FEBS Letters* 2010, **584**:3682–3695.
2. Stracker TH, Petrini JHJ: **The MRE11 complex: starting from the ends.** *Nature Publishing Group* 2011, **12**:90–103.
3. Bhatti S, Kozlov S, Farooqi AA, Naqi A, Lavin M, Khanna KK: **ATM protein kinase: the linchpin of cellular defenses to stress.** *Cell Mol Life Sci* 2011, **68**:2977–3006.
4. Fischle W: **Tip60-ing the balance in DSB repair.** *Nat Cell Biol* 2009, **11**:1279–1281.
5. Sun Y, Jiang X, Chen S, Fernandes N, Price BD: **A role for the Tip60 histone acetyltransferase in the acetylation and activation of ATM.** *Proc Natl Acad Sci USA* 2005, **102**:13182–13187.
6. Derheimer FA, Kastan MB: **Multiple roles of ATM in monitoring and maintaining DNA**

integrity. *FEBS Letters* 2010, **584**:3675–3681.
7. Bakkenist CJ, Kastan MB: **DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation.** *Nature* 2003, **421**:499–506.
8. Kozlov SV, Graham ME, Jakob B, Tobias F, Kijas AW, Tanuji M, Chen P, Robinson PJ, Taucher-Scholz G, Suzuki K, So S, Chen D, Lavin MF: **Autophosphorylation and ATM Activation: ADDITIONAL SITES ADD TO THE COMPLEXITY**. *J Biol Chem* 2011, **286**:9107–9119.
9. Yuan J, Adamski R, Chen J: **Focus on histone variant H2AX: To be or not to be**. *FEBS Letters* 2010, **584**:3717–3724.
10. Chapman JR, Jackson SP: **Phospho-dependent interactions between NBS1 and MDC1 mediate chromatin retention of the MRN complex at sites of DNA damage.** *EMBO reports* 2008, **9**:795–801.
11. Spycher C, Miller ES, Townsend K, Pavic L, Morrice NA, Janscak P, Stewart GS, Stucki M: **Constitutive phosphorylation of MDC1 physically links the MRE11-RAD50-NBS1 complex to damaged chromatin.** *The Journal of Cell Biology* 2008, **181**:227–240.
12. van Attikum H, Gasser SM: **Crosstalk between histone modifications during the DNA damage response.** *Trends in Cell Biology* 2009, **19**:207–217.
13. Bao Y: **Chromatin response to DNA double-strand break damage.** *Epigenomics* 2011, **3**:307–321.
14. Campbell SJ, Edwards RA, Leung CCY, Neculai D, Hodge CD, Dhe-Paganon S, Glover JNM: **Molecular insights into the function of RING finger (RNF)-containing proteins hRNF8 and hRNF168 in Ubc13/Mms2-dependent ubiquitylation.** *J Biol Chem* 2012, **287**:23900–23910.
15. Bekker-Jensen S, Rendtlew Danielsen J, Fugger K, Gromova I, Nerstedt A, Lukas C, Bartek J, Lukas J, Mailand N: **HERC2 coordinates ubiquitin-dependent assembly of DNA repair factors on damaged chromosomes.** *Nat Cell Biol* 2010, **12**:80–6– sup pp 1–12.
16. Doil C, Mailand N, Bekker-Jensen S, Menard P, Larsen DH, Pepperkok R, Ellenberg J, Panier S, Durocher D, Bartek J, Lukas J, Lukas C: **RNF168 Binds and Amplifies Ubiquitin Conjugates on Damaged Chromosomes to Allow Accumulation of Repair Proteins**. *Cell* 2009, **136**:435–446.
17. Stewart GS, Panier S, Townsend K, Al-Hakim AK, Kolas NK, Miller ES, Nakada S, Ylanko J, Olivarius S, Mendez M, Oldreive C, in JW, Tagliaferro A, Pelletier L, Taubenheim N, Durandy A, Byrd PJ, Stankovic T, Taylor AMR, Durocher D: **The RIDDLE Syndrome Protein Mediates a Ubiquitin-Dependent Signaling Cascade at Sites of DNA Damage**. *Cell* 2009, **136**:420–434.
18. Larsen DH, Poinsignon C, Gudjonsson T, Dinant C, Payne MR, Hari FJ, Rendtlew Danielsen JM, Menard P, Sand JC, Stucki M, Lukas C, Bartek J, Andersen JS, Lukas J: **The chromatin-remodeling factor CHD4 coordinates signaling and repair after DNA damage**. *The Journal of Cell Biology* 2010, **190**:731–740.
19. Gudjonsson T, Altmeyer M, Savic V, Toledo L, Dinant C, Grøfte M, Bartkova J, Poulsen M, Oka Y, Bekker-Jensen S, Mailand N, Neumann B, Heriche J-K, Shearer R, Saunders D, Bartek J, Lukas J, Lukas C: **TRIP12 and UBR5 Suppress Spreading of Chromatin Ubiquitylation at Damaged Chromosomes**. *Cell* 2012, **150**:697–709.
20. Nakada S, Tai I, Panier S, Al-Hakim A, Iemura S-I, Juang Y-C, O'Donnell L, Kumakubo A, Munro M, Sicheri F, Gingras A-C, Natsume T, Suda T, Durocher D: **Non-canonical inhibition of DNA damage-dependent ubiquitination by OTUB1**. *Nature* 2010, **466**:941–946.
21. Roy R, Chun J, Powell SN: **BRCA1 and BRCA2: different roles in a common pathway of genome protection.** *Nat Rev Cancer* 2012, **12**:68–78.
22. Huen MSY, Sy SMH, Chen J: **BRCA1 and its toolbox for the maintenance of genome integrity.** *Multiple values selected* 2010, **11**:138–148.
23. Ohta T, Sato K, Wu W: **The BRCA1 ubiquitin ligase and homologous recombination repair**. *FEBS Letters* 2011, **585**:2836–2844.

24. Bouwman P, Aly A, Escandell JM, Pieterse M, Bartkova J, van der Gulden H, Hiddingh S, Thanasoula M, Kulkarni A, Yang Q, Haffty BG, Tommiska J, Blomqvist C, Drapkin R, Adams DJ, Nevanlinna H, Bartek J, Tarsounas M, Ganesan S, Jonkers J: **53BP1 loss rescues BRCA1 deficiency and is associated with triple-negative and BRCA-mutated breast cancers.** *Nature Structural & Molecular Biology* 2010, **17**:688–695.
25. Aly A, Ganesan S: **BRCA1, PARP, and 53BP1: conditional synthetic lethality and synthetic viability**. *Journal of Molecular Cell Biology* 2011, **3**:66–74.
26. Buis J, Stoneham T, Spehalski E, Ferguson DO: **Mre11 regulates CtIP-dependent double-strand break repair by interaction with CDK2**. *Nature Structural & Molecular Biology* 2012, **19**:246–252.
27. Huertas P, Jackson SP: **Human CtIP mediates cell cycle control of DNA end resection and double strand break repair.** *J Biol Chem* 2009, **284**:9558–9565.
28. Barber LJ, Boulton SJ: **BRCA1 ubiquitylation of CtIP: Just the tIP of the iceberg?** *DNA Repair* 2006, **5**:1499–1504.
29. Chen L, Nievera CJ, Lee AY-L, Wu X: **Cell cycle-dependent complex formation of BRCA1.CtIP.MRN is important for DNA double-strand break repair.** *J Biol Chem* 2008, **283**:7713–7720.
30. Symington LS, Gautier J: **Double-Strand Break End Resection and Repair Pathway Choice**. *Annu Rev Genet* 2011, **45**:247–271.
31. Bernstein KA, Rothstein R: **At Loose Ends: Resecting a Double-Strand Break**. *Cell* 2009, **137**:807–810.
32. Tomimatsu N, Mukherjee B, Deland K, Kurimasa A, Bolderson E, Khanna KK, Burma S: **Exo1 plays a major role in DNA end resection in humans and influences double-strand break repair and damage signaling decisions.** *DNA Repair* 2012, **11**:441–448.
33. Bolderson E, Tomimatsu N, Richard DJ, Boucher D, Kumar R, Pandita TK, Burma S, Khanna KK: **Phosphorylation of Exo1 modulates homologous recombination repair of DNA double-strand breaks.** *Nucleic Acids Research* 2010, **38**:1821–1831.
34. Nimonkar AV, Genschel J, Kinoshita E, Polaczek P, Campbell JL, Wyman C, Modrich P, Kowalczykowski SC: **BLM-DNA2-RPA-MRN and EXO1-BLM-RPA-MRN constitute two DNA end resection machineries for human DNA break repair**. *Genes & Development* 2011, **25**:350–362.
35. Karanja KK, Cox SW, Duxin JP, Stewart SA, Campbell JL: **DNA2 and EXO1 in replication-coupled, homology-directed repair and in the interplay between HDR and the FA/BRCA network**. *Cell Cycle* 2012, **11**:3983–3996.
36. San Filippo J, Sung P, Klein H: **Mechanism of Eukaryotic Homologous Recombination**. *Annu Rev Biochem* 2008, **77**:229–257.
37. Moynahan ME, Jasin M: **Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis.** *Nature Publishing Group* 2010, **11**:196–207.
38. Ciccia A, Elledge SJ: **The DNA damage response: making it safe to play with knives.** *Molecular Cell* 2010, **40**:179–204.
39. Dong Y, Hakimi M-A, Chen X, Kumaraswamy E, Cooch NS, Godwin AK, Shiekhattar R: **Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in DNA repair.** *Molecular Cell* 2003, **12**:1087–1099.
40. Suwaki N, Klare K, Tarsounas M: **RAD51 paralogs: Roles in DNA damage signalling, recombinational repair and tumorigenesis**. *Seminars in Cell and Developmental Biology* 2011, **22**:898–905.
41. Feng Z, Scott SP, Bussen W, Sharma GG, Guo G, Pandita TK, Powell SN: **Rad52 inactivation is synthetically lethal with BRCA2 deficiency.** *Proceedings of the National Academy of Sciences* 2011, **108**:686–691.
42. Mazin AV, Mazina OM, Bugreev DV, Rossi MJ: **Rad54, the motor of homologous recombination.** *DNA Repair* 2010, **9**:286–302.
43. Ceballos SJ, Heyer W-D: **Functions of the Snf2/Swi2 family Rad54 motor protein in**

homologous recombination. *Biochim Biophys Acta* 2011, **1809**:509–523.
44. Modesti M, Budzowska M, Baldeyron C, Demmers JAA, Ghirlando R, Kanaar R: **RAD51AP1 is a structure-specific DNA binding protein that stimulates joint molecule formation during RAD51-mediated homologous recombination.** *Molecular Cell* 2007, **28**:468–481.
45. Wiese C, Dray E, Groesser T, San Filippo J, Shi I, Collins DW, Tsai M-S, Williams GJ, Rydberg B, Sung P, Schild D: **Promotion of homologous recombination and genomic stability by RAD51AP1 via RAD51 recombinase enhancement.** *Molecular Cell* 2007, **28**:482–490.
46. Richard DJ, Bolderson E, Cubeddu L, Wadsworth RIM, Savage K, Sharma GG, Nicolette ML, Tsvetanov S, McIlwraith MJ, Pandita RK, Takeda S, Hay RT, Gautier J, West SC, Paull TT, Pandita TK, White MF, Khanna KK: **Single-stranded DNA-binding protein hSSB1 is critical for genomic stability.** *Nature* 2008, **453**:677–681.
47. Li Y, Bolderson E, Kumar R, Muniandy PA, Xue Y, Richard DJ, Seidman M, Pandita TK, Khanna KK, Wang W: **HSSB1 and hSSB2 form similar multiprotein complexes that participate in DNA damage response.** *J Biol Chem* 2009, **284**:23525–23531.
48. Barber LJ, Youds JL, Ward JD, McIlwraith MJ, O'Neil NJ, Petalcorin MIR, Martin JS, Collis SJ, Cantor SB, Auclair M, Tissenbaum H, West SC, Rose AM, Boulton SJ: **RTEL1 maintains genomic stability by suppressing homologous recombination.** *Cell* 2008, **135**:261–271.
49. Moldovan G-L, Dejsuphong D, Petalcorin MIR, Hofmann K, Takeda S, Boulton SJ, D'Andrea AD: **Inhibition of homologous recombination by the PCNA-interacting protein PARI.** *Molecular Cell* 2012, **45**:75–86.
50. McIlwraith MJ, Vaisman A, Liu Y, Fanning E, Woodgate R, West SC: **Human DNA Polymerase η Promotes DNA Synthesis from Strand Invasion Intermediates of Homologous Recombination**. *Molecular Cell* 2005, **20**:783–792.
51. McIlwraith MJ, West SC: **DNA Repair Synthesis Facilitates RAD52-Mediated Second-End Capture during DSB Repair**. *Molecular Cell* 2008, **29**:510–516.
52. Mazón G, Mimitou EP, Symington LS: **SnapShot: Homologous Recombination in DNA Double-Strand Break Repair**. *Cell* 2010, **142**:648.e1–648.e2.
53. Maloisel L, Fabre F, Gangloff S: **DNA polymerase delta is preferentially recruited during homologous recombination to promote heteroduplex DNA extension**. *Molecular and Cellular Biology* 2008, **28**:1373–1382.
54. Chang JH, Kim JJ, Choi JM, Lee JH, Cho Y: **Crystal structure of the Mus81-Eme1 complex.** *Genes & Development* 2008, **22**:1093–1106.
55. Ciccia A, McDonald N, West SC: **Structural and functional relationships of the XPF/MUS81 family of proteins.** *Annu Rev Biochem* 2008, **77**:259–287.
56. Xu D, Guo R, Sobeck A, Bachrati CZ, Yang J, Enomoto T, Brown GW, Hoatlin ME, Hickson ID, Wang W: **RMI, a new OB-fold complex essential for Bloom syndrome protein to maintain genome stability.** *Genes & Development* 2008, **22**:2843–2855.
57. Rass U, Compton SA, Matos J, Singleton MR, Ip SCY, Blanco MG, Griffith JD, West SC: **Mechanism of Holliday junction resolution by the human GEN1 protein.** *Genes & Development* 2010, **24**:1559–1569.
58. Fekairi S, Scaglione S, Chahwan C, Taylor ER, Tissier A, Coulon S, Dong M-Q, Ruse C, Yates JR, Russell P, Fuchs RP, McGowan CH, Gaillard P-HL: **Human SLX4 is a Holliday junction resolvase subunit that binds multiple DNA repair/recombination endonucleases.** *Cell* 2009, **138**:78–89.
59. Grimme JM, Honda M, Wright R, Okuno Y, Rothenberg E, Mazin AV, Ha T, Spies M: **Human Rad52 binds and wraps single-stranded DNA and mediates annealing via two hRad52-ssDNA complexes.** *Nucleic Acids Research* 2010, **38**:2917–2930.
60. Lok BH, Powell SN: **Molecular pathways: understanding the role of Rad52 in homologous recombination for therapeutic advancement.** *Clin Cancer Res* 2012, **18**:6400–6406.
61. McNeil EM, Melton DW: **DNA repair endonuclease ERCC1-XPF as a novel therapeutic target to overcome chemoresistance in cancer therapy.** *Nucleic Acids Research* 2012, **40**:9990–10004.

62. Tripsianes K, Folkers G, AB E, Das D, Odijk H, Jaspers NGJ, Hoeijmakers JHJ, Kaptein R, Boelens R: **The Structure of the Human ERCC1/XPF Interaction Domains Reveals a Complementary Role for the Two Proteins in Nucleotide Excision Repair**. *Structure* 2005, **13**:1849–1858.
63. Göttlich B, Reichenberger S, Feldmann E, Pfeiffer P: **Rejoining of DNA double-strand breaks in vitro by single-strand annealing.** *Eur J Biochem* 1998, **258**:387–395.
64. Galanty Y, Belotserkovskaya R, Coates J, Polo S, Miller KM, Jackson SP: **Mammalian SUMO E3-ligases PIAS1 and PIAS4 promote responses to DNA double-strand breaks.** *Nature* 2009, **462**:935–939.
65. Morris JR, Boutell C, Keppler M, Densham R, Weekes D, Alamshah A, Butler L, Galanty Y, Pangon L, Kiuchi T, Ng T, Solomon E: **The SUMO modification pathway is involved in the BRCA1 response to genotoxic stress.** *Nature* 2009, **462**:886–890.
66. Adamson B, Smogorzewska A, Sigoillot FD, King RW, Elledge SJ: **A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response.** *Nat Cell Biol* 2012, **14**:318–328.
67. Li J, Zou C, Bai Y, Wazer DE, Band V, Gao Q: **DSS1 is required for the stability of BRCA2.** *Oncogene* 2006, **25**:1186–1194.
68. Bekker-Jensen S, Mailand N: **The ubiquitin- and SUMO-dependent signaling response to DNA double-strand breaks.** *FEBS Letters* 2011, **585**:2914–2919.
69. Danielsen JR, Povlsen LK, Villumsen BH, Streicher W, Nilsson J, Wikström M, Bekker-Jensen S, Mailand N: **DNA damage-inducible SUMOylation of HERC2 promotes RNF8 binding via a novel SUMO-binding Zinc finger.** *The Journal of Cell Biology* 2012, **197**:179–187.
70. Shiloh Y, Shema E, Moyal L, Oren M: **RNF20-RNF40: A ubiquitin-driven link between gene expression and the DNA damage response.** *FEBS Letters* 2011, **585**:2795–2802.
71. Nakamura K, Kato A, Kobayashi J, Yanagihara H, Sakamoto S, Oliveira DVNP, Shimada M, Tauchi H, Suzuki H, Tashiro S, Zou L, Komatsu K: **Regulation of homologous recombination by RNF20-dependent H2B ubiquitination.** *Molecular Cell* 2011, **41**:515–528.
72. Moyal L, Lerenthal Y, Gana-Weisz M, Mass G, So S, Wang S-Y, Eppink B, Chung YM, Shalev G, Shema E, Shkedy D, Smorodinsky NI, van Vliet N, Kuster B, Mann M, Ciechanover A, Dahm-Daphi J, Kanaar R, Hu MCT, Chen DJ, Oren M, Shiloh Y: **Requirement of ATM-dependent monoubiquitylation of histone H2B for timely repair of DNA double-strand breaks.** *Molecular Cell* 2011, **41**:529–542.
73. Galanty Y, Belotserkovskaya R, Coates J, Jackson SP: **RNF4, a SUMO-targeted ubiquitin E3 ligase, promotes DNA double-strand break repair.** *Genes & Development* 2012, **26**:1179–1195.
74. Yin Y, Seifert A, Chua JS, Maure J-F, Golebiowski F, Hay RT: **SUMO-targeted ubiquitin E3 ligase RNF4 is required for the response of human cells to DNA damage.** *Genes & Development* 2012, **26**:1196–1208.
75. Luo K, Zhang H, Wang L, Yuan J, Lou Z: **Sumoylation of MDC1 is important for proper DNA damage response.** *EMBO J* 2012, **31**:3008–3019.
76. Lieber MR: **The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway.** *Annu Rev Biochem* 2010, **79**:181–211.
77. Lieber MR: **The mechanism of human nonhomologous DNA end joining.** *J Biol Chem* 2008, **283**:1–5.
78. Lieber MR: **NHEJ and its backup pathways in chromosomal translocations**. *Nature Structural & Molecular Biology* 2010, **17**:393–395.
79. Bothmer A, Robbiani DF, Feldhahn N, Gazumyan A, Nussenzweig A, Nussenzweig MC: **53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination.** *Journal of Experimental Medicine* 2010, **207**:855–865.
80. Wang M, Wu W, Wu W, Rosidi B, Zhang L, Wang H, Iliakis G: **PARP-1 and Ku compete for repair of DNA double strand breaks by distinct NHEJ pathways.** *Nucleic Acids Research*

2006, **34**:6170–6182.
81. Patel AG, Sarkaria JN, Kaufmann SH: **Nonhomologous end joining drives poly(ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells.** *Proceedings of the National Academy of Sciences* 2011, **108**:3406–3411.
82. Roberts SA, Strande N, Burkhalter MD, Strom C, Havener JM, Hasty P, Ramsden DA: **Ku is a 5'-dRP/AP lyase that excises nucleotide damage near broken ends.** *Nature* 2010, **464**:1214–1217.
83. Mahaney BL, Meek K, Lees-Miller SP: **Repair of ionizing radiation-induced DNA double-strand breaks by non-homologous end-joining.** *Biochem J* 2009, **417**:639–650.
84. Dobbs TA, Tainer JA, Lees-Miller SP: **A structural model for regulation of NHEJ by DNA-PKcs autophosphorylation**. *DNA Repair* 2010, **9**:1307–1314.
85. Yannone SM, Khan IS, Zhou RZ, Zhou T, Valerie K, Povirk LF: **Coordinate 5“ and 3” endonucleolytic trimming of terminally blocked blunt DNA double-strand break ends by Artemis nuclease and DNA-dependent protein kinase**. *Nucleic Acids Research* 2008, **36**:3354–3365.
86. Lieber MR, Gu J, Lu H, Shimazaki N, Tsai AG: **Nonhomologous DNA end joining (NHEJ) and chromosomal translocations in humans.** *Subcell Biochem* 2010, **50**:279–296.
87. Drouet J: **DNA-dependent Protein Kinase and XRCC4-DNA Ligase IV Mobilization in the Cell in Response to DNA Double Strand Breaks**. *Journal of Biological Chemistry* 2005, **280**:7060–7069.
88. Wang Y, Lamarche BJ, Tsai M-D: **Human DNA Ligase IV and the Ligase IV/XRCC4 Complex: Analysis of Nick Ligation Fidelity †**. *Biochemistry* 2007, **46**:4962–4976.
89. Ahnesorg P, Smith P, Jackson SP: **XLF Interacts with the XRCC4-DNA Ligase IV Complex to Promote DNA Nonhomologous End-Joining**. *Cell* 2006, **124**:301–313.
90. Nick McElhinny SA, Snowden CM, McCarville J, Ramsden DA: **Ku recruits the XRCC4-ligase IV complex to DNA ends.** *Molecular and Cellular Biology* 2000, **20**:2996–3003.
91. Yano K-I, Morotomi-Yano K, Wang S-Y, Uematsu N, Lee K-J, Asaithamby A, Weterings E, Chen DJ: **Ku recruits XLF to DNA double-strand breaks**. *EMBO reports* 2007, **9**:91–96.
92. Cheng Q, Barboule N, Frit P, Gomez D, Bombarde O, Couderc B, Ren GS, Salles B, Calsou P: **Ku counteracts mobilization of PARP1 and MRN in chromatin damaged with DNA double-strand breaks**. *Nucleic Acids Research* 2011, **39**:9605–9619.
93. Lieber MR: **NHEJ and its backup pathways in chromosomal translocations**. *Nature Structural & Molecular Biology* 2010, **17**:393–395.
94. Rupnik A, Lowndes NF, Grenon M: **MRN and the race to the break**. *Chromosoma* 2009, **119**:115–135.
95. You Z, Shi LZ, Zhu Q, Wu P, Zhang Y-W, Basilio A, Tonnu N, Verma IM, Berns MW, Hunter T: **CtIP Links DNA Double-Strand Break Sensing to Resection**. *Molecular Cell* 2009, **36**:954–969.
96. Audebert M: **Involvement of Poly(ADP-ribose) Polymerase-1 and XRCC1/DNA Ligase III in an Alternative Route for DNA Double-strand Breaks Rejoining**. *Journal of Biological Chemistry* 2004, **279**:55117–55126.
97. Wang H, Rosidi B, Perrault R, Wang M, Zhang L, Windhofer F, Iliakis G: **DNA ligase III as a candidate component of backup pathways of nonhomologous end joining.** *Cancer Research* 2005, **65**:4020–4030.
98. Mladenov E, Iliakis G: **Induction and repair of DNA double strand breaks: The increasing spectrum of non-homologous end joining pathways**. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 2011, **711**:61–72.
99. Sertic S, Pizzi S, Lazzaro F, Plevani P, Muzi-Falconi M: **NER and DDR: Classical music with new instruments**. *Cell Cycle* 2012, **11**:668–674.
100. Kamileri I, Karakasilioti I, Garinis GA: **Nucleotide excision repair: new tricks with old bricks**. *Trends in Genetics* 2012, **28**:566–573.
101. Nishi R, Okuda Y, Watanabe E, Mori T, Iwai S, Masutani C, Sugasawa K, Hanaoka F: **Centrin**

2 Stimulates Nucleotide Excision Repair by Interacting with Xeroderma Pigmentosum Group C Protein. *Molecular and Cellular Biology* 2005, **25**:5664–5674.
102. Nouspikel T: **DNA repair in mammalian cells : Nucleotide excision repair: variations on versatility.** *Cell Mol Life Sci* 2009, **66**:994–1009.
103. Compe E, Egly J-M: **TFIIH: when transcription met DNA repair**. *Nature Reviews Molecular Cell Biology* 2012, **13**:343–354.
104. Saijo M, Takedachi A, Tanaka K: **Nucleotide excision repair by mutant xeroderma pigmentosum group A (XPA) proteins with deficiency in interaction with RPA.** *J Biol Chem* 2011, **286**:5476–5483.
105. Sugasawa K: **Regulation of damage recognition in mammalian global genomic nucleotide excision repair**. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* 2010, **685**:29–37.
106. Zotter A, Luijsterburg MS, Warmerdam DO, Ibrahim S, Nigg A, van Cappellen WA, Hoeijmakers JHJ, van Driel R, Vermeulen W, Houtsmuller AB: **Recruitment of the Nucleotide Excision Repair Endonuclease XPG to Sites of UV-Induced DNA Damage Depends on Functional TFIIH**. *Molecular and Cellular Biology* 2006, **26**:8868–8879.
107. Caldecott KW, McKeown CK, Tucker JD, Ljungquist S, Thompson LH: **An interaction between the mammalian DNA repair protein XRCC1 and DNA ligase III.** *Molecular and Cellular Biology* 1994, **14**:68–76.
108. Mocquet V, Lainé JP, Riedl T, Yajin Z, Lee MY, Egly J-M: **Sequential recruitment of the repair factors during NER: the role of XPG in initiating the resynthesis step**. *EMBO J* 2007, **27**:155–167.
109. Moser J, Kool H, Giakzidis I, Caldecott K, Mullenders LHF, Fousteri MI: **Sealing of Chromosomal DNA Nicks during Nucleotide Excision Repair Requires XRCC1 and DNA Ligase IIIα in a Cell-Cycle-Specific Manner**. *Molecular Cell* 2007, **27**:311–323.
110. Lainé JP, Egly J-M: **Initiation of DNA repair mediated by a stalled RNA polymerase IIO.** *EMBO J* 2006, **25**:387–397.
111. Hanawalt PC, Spivak G: **Transcription-coupled DNA repair: two decades of progress and surprises**. *Nature Reviews Molecular Cell Biology* 2008, **9**:958–970.
112. Lagerwerf S, Vrouwe MG, Overmeer RM, Fousteri MI, Mullenders LHF: **DNA damage response and transcription**. *DNA Repair* 2011, **10**:743–750.
113. Dianov GL: **Base excision repair targets for cancer therapy.** *Am J Cancer Res* 2011, **1**:845–851.
114. Liu Y, Wilson SH: **DNA base excision repair: a mechanism of trinucleotide repeat expansion**. *Trends in Biochemical Sciences* 2012, **37**:162–172.
115. Jacobs AL, Schär P: **DNA glycosylases: in DNA repair and beyond**. *Chromosoma* 2011, **121**:1–20.
116. Maynard S, Schurman SH, Harboe C, de Souza-Pinto NC, Bohr VA: **Base excision repair of oxidative DNA damage and association with cancer and aging**. *Carcinogenesis* 2008, **30**:2–10.
117. Asagoshi K, Liu Y, Masaoka A, Lan L, Prasad R, Horton JK, Brown AR, Wang X-H, Bdour HM, Sobol RW: **DNA polymerase β-dependent long patch base excision repair in living cells**. *DNA Repair* 2010, **9**:109–119.
118. Robertson AB, Klungland A, Rognes T, Leiros I: **DNA repair in mammalian cells: Base excision repair: the long and short of it.** *Cell Mol Life Sci* 2009, **66**:981–993.
119. Dianova II: **XRCC1-DNA polymerase interaction is required for efficient base excision repair**. *Nucleic Acids Research* 2004, **32**:2550–2555.
120. Parsons JL, Dianova II, Allinson SL, Dianov GL: **DNA Polymerase β Promotes Recruitment of DNA Ligase IIIα−XRCC1 to Sites of Base Excision Repair**. *Biochemistry* 2005, **44**:10613–10619.
121. Lange SS, Takata K-I, Wood RD: **DNA polymerases and cancer.** *Nat Rev Cancer* 2011, **11**:96–110.

122. Asagoshi K, Tano K, Chastain PD, Adachi N, Sonoda E, Kikuchi K, Koyama H, Nagata K, Kaufman DG, Takeda S, Wilson SH, Watanabe M, Swenberg JA, Nakamura J: **FEN1 Functions in Long Patch Base Excision Repair Under Conditions of Oxidative Stress in Vertebrate Cells**. *Molecular Cancer Research* 2010, **8**:204–215.
123. Mukherjee S, Ridgeway AD, Lamb DJ: **DNA mismatch repair and infertility**. *Current Opinion in Urology* 2010, **20**:525–532.
124. Hsieh P, Yamane K: **DNA mismatch repair: Molecular mechanism, cancer, and ageing**. *Mechanisms of Ageing and Development* 2008, **129**:391–407.
125. Peña-Diaz J, Jiricny J: **Mammalian mismatch repair: error-free or error-prone?** *Trends in Biochemical Sciences* 2012, **37**:206–214.
126. Drotschmann K: **Asymmetric Recognition of DNA local distortion. Structure-based functional studies of eukaryotic Msh2-Msh6**. *Journal of Biological Chemistry* 2001, **276**:46225–46229.
127. Martin SA, Lord CJ, Ashworth A: **Therapeutic Targeting of the DNA Mismatch Repair Pathway**. *Clinical Cancer Research* 2010, **16**:5107–5113.
128. Moldovan G-L, Pfander B, Jentsch S: **PCNA, the Maestro of the Replication Fork**. *Cell* 2007, **129**:665–679.
129. Jiricny J: **The multifaceted mismatch-repair system**. *Nature Reviews Molecular Cell Biology* 2006, **7**:335–346.
130. Moldovan G-L, D'Andrea AD: **How the Fanconi Anemia Pathway Guards the Genome**. *Annu Rev Genet* 2009, **43**:223–249.
131. Kim H, D'Andrea AD: **Regulation of DNA cross-link repair by the Fanconi anemia/BRCA pathway**. *Genes & Development* 2012, **26**:1393–1408.
132. Singh TR, Saro D, Ali AM, Zheng X-F, Du C-H, Killen MW, Sachpatzidis A, Wahengbam K, Pierce AJ, Xiong Y, Sung P, Meetei AR: **MHF1-MHF2, a Histone-Fold-Containing Protein Complex, Participates in the Fanconi Anemia Pathway via FANCM**. *Molecular Cell* 2010, **37**:879–886.
133. Yan Z, Delannoy M, Ling C, Daee D, Osman F, Muniandy PA, Shen X, Oostra AB, Du H, Steltenpool J, Lin T, Schuster B, DEcaillet C, Stasiak A, Stasiak AZ, Stone S, Hoatlin ME, Schindler D, Woodcock CL, Joenje H, Sen R, de Winter JP, Li L, Seidman MM, Whitby MC, Myung K, Constantinou A, Wang W: **A Histone-Fold Complex and FANCM Form a Conserved DNA-Remodeling Complex to Maintain Genome Stability**. *Molecular Cell* 2010, **37**:865–878.
134. Crossan GP, Patel KJ: **The Fanconi anaemia pathway orchestrates incisions at sites of crosslinked DNA**. *J Pathol* 2011, **226**:326–337.
135. Jensen SRSRA, Jensen RA: **Identifying and exploiting defects in the Fanconi anemia/BRCA pathway in oncology**. *Translational Research* 2012, **160**:178–197.
136. Kim JM, Kee Y, Gurtan A, D'Andrea AD: **Cell cycle-dependent chromatin loading of the Fanconi anemia core complex by FANCM/FAAP24**. *Blood* 2008, **111**:5215–5222.
137. Smogorzewska A, Matsuoka S, Vinciguerra P, McDonald ER III, Hurov KE, Luo J, Ballif BA, Gygi SP, Hofmann K, D'Andrea AD, Elledge SJ: **Identification of the FANCI Protein, a Monoubiquitinated FANCD2 Paralog Required for DNA Repair**. *Cell* 2007, **129**:289–301.
138. Kee Y, D'Andrea AD: **Expanded roles of the Fanconi anemia pathway in preserving genomic stability**. *Genes & Development* 2010, **24**:1680–1694.
139. Garner E, Smogorzewska A: **Ubiquitylation and the Fanconi anemia pathway**. *FEBS Letters* 2011, **585**:2853–2860.
140. D'Andrea AD: **Susceptibility pathways in Fanconi's anemia and breast cancer.** *N Engl J Med* 2010, **362**:1909–1919.
141. Cybulski KE, Howlett NG: **FANCP/SLX4: a Swiss army knife of DNA interstrand crosslink repair.** *Cell Cycle* 2011, **10**:1757–1763.
142. Gan GN, Wittschieben JP, Wittschieben BØ, Wood RD: **DNA polymerase zeta (pol ζ) in higher eukaryotes**. *Cell Res* 2008, **18**:174–183.

143. Shen X: **REV3 and REV1 Play Major Roles in Recombination-independent Repair of DNA Interstrand Cross-links Mediated by Monoubiquitinated Proliferating Cell Nuclear Antigen (PCNA)**. *Journal of Biological Chemistry* 2006, **281**:13869–13872.
144. Deans AJ, West SC: **DNA interstrand crosslink repair and cancer.** *Nat Rev Cancer* 2011, **11**:467–480.
145. Nijman SMB, Huang TT, Dirac AMG, Brummelkamp TR, Kerkhoven RM, D'Andrea AD, Bernards R: **The Deubiquitinating Enzyme USP1 Regulates the Fanconi Anemia Pathway**. *Molecular Cell* 2005, **17**:331–339.
146. Cohn MA, Kowal P, Yang K, Haas W, Huang TT, Gygi SP, D'Andrea AD: **A UAF1-Containing Multisubunit Protein Complex Regulates the Fanconi Anemia Pathway**. *Molecular Cell* 2007, **28**:786–797.

Appendix Two: Supplementary figures of Chapter Three*

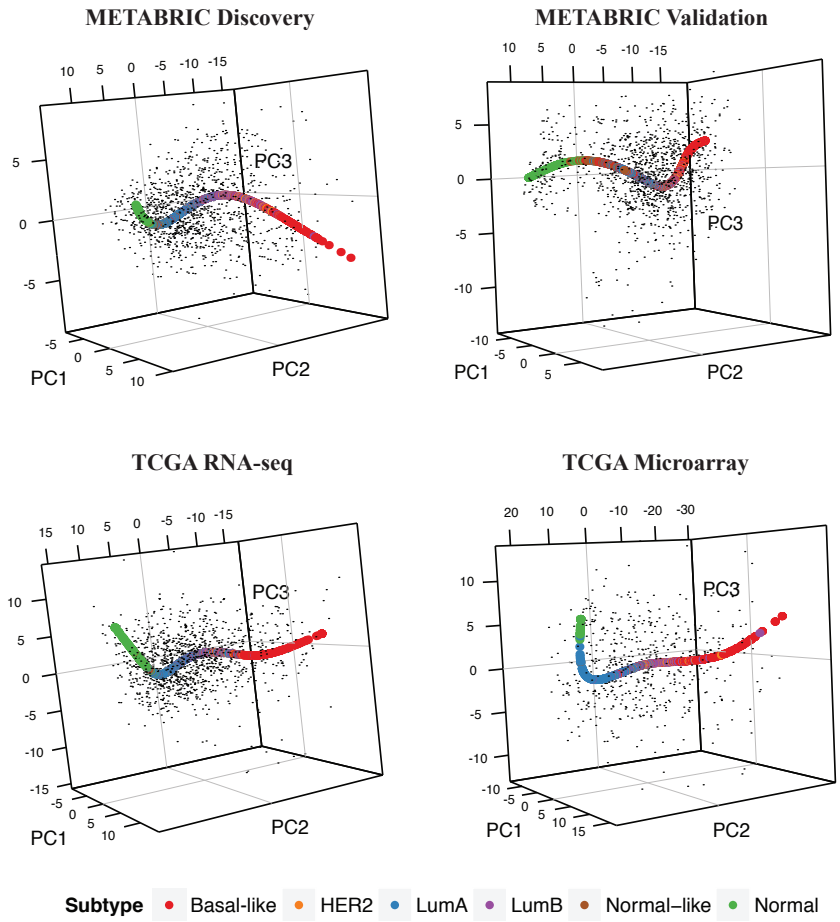


Figure 1 Principal curve of the HR pathway for each of the four cohorts. For each cohort, the black points represent samples in that cohort. The samples are projected onto the principal curve and are coloured according to their PAM50 assignment. The data points and the principal curve are projected on the three leading principal components for visualisation.[1]

*The supplementary tables of Chapter Three, which are all Excel files, are deposited at UQ eSpace.

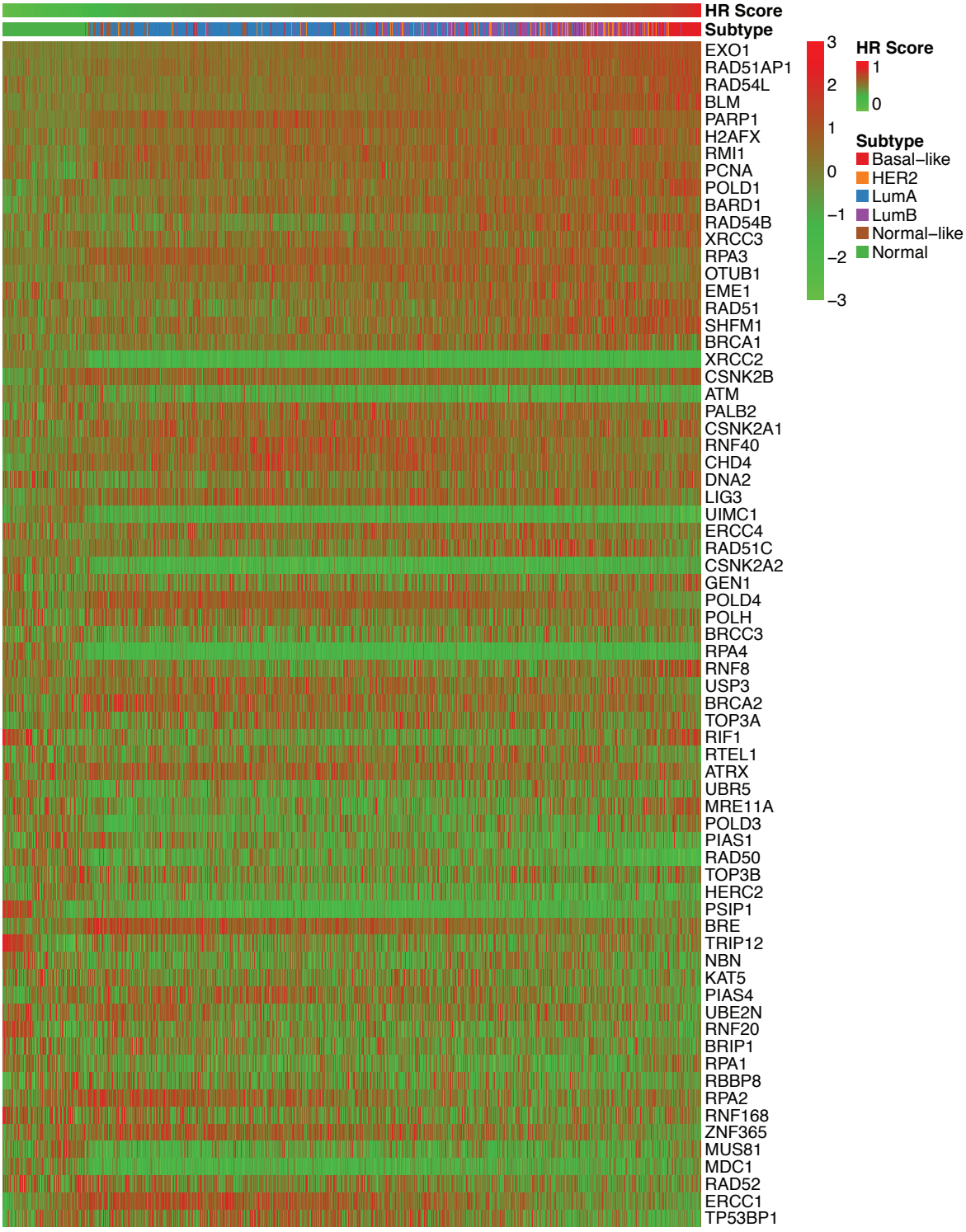


Figure 2 Expression of the HR genes in tumours from the METABRIC Discovery cohort. The HR genes are ranked in decreasing importance according to their contribution to the first principal component.

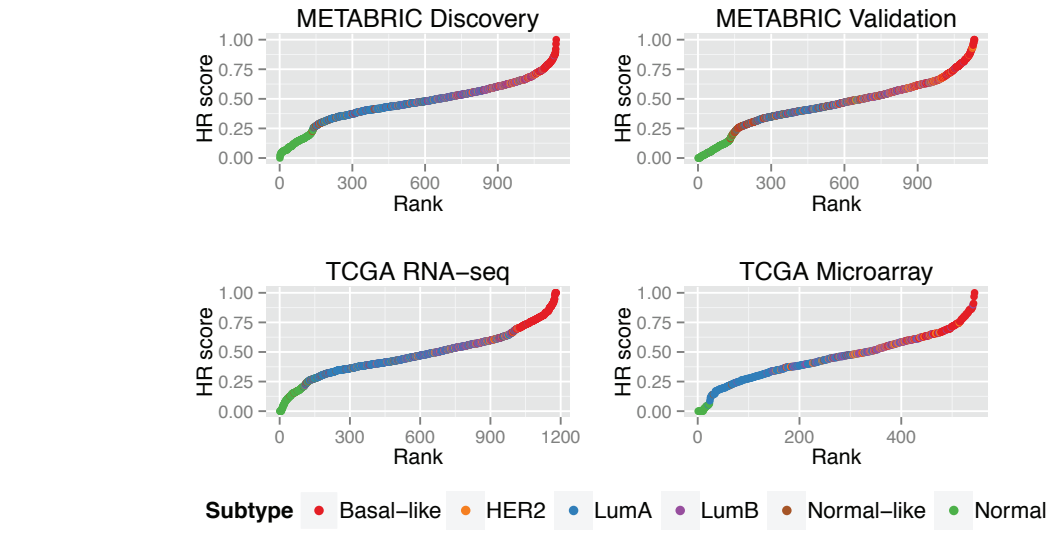


Figure 3 Scatter plots of the *HR* score versus the rank of tumours according to their *HR* score, colour by the PAM50 assignment.

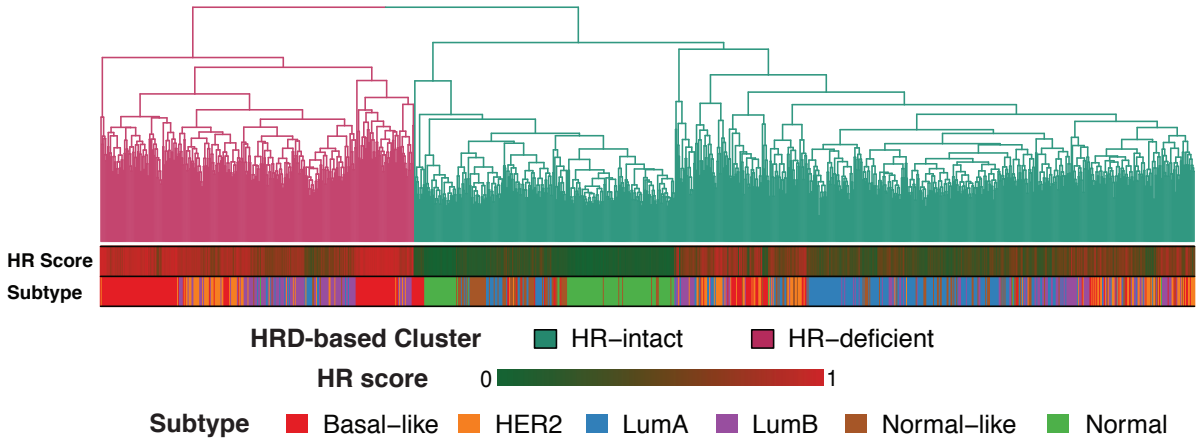


Figure 4 Hierarchical clustering of tumours from the METABRIC Validation cohort based on the HRD signature.

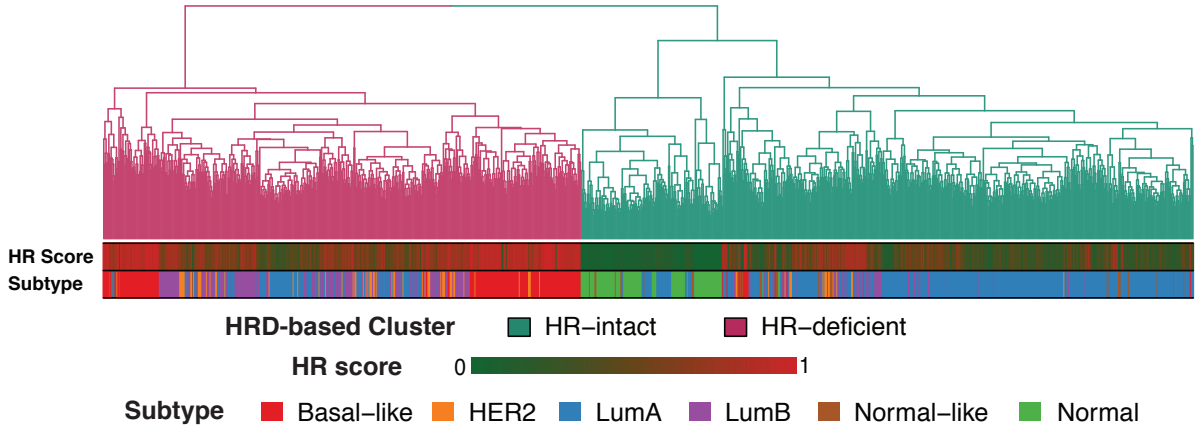


Figure 5 Hierarchical clustering of tumours from the TCGA RNA-seq cohort based on the HRD signature.

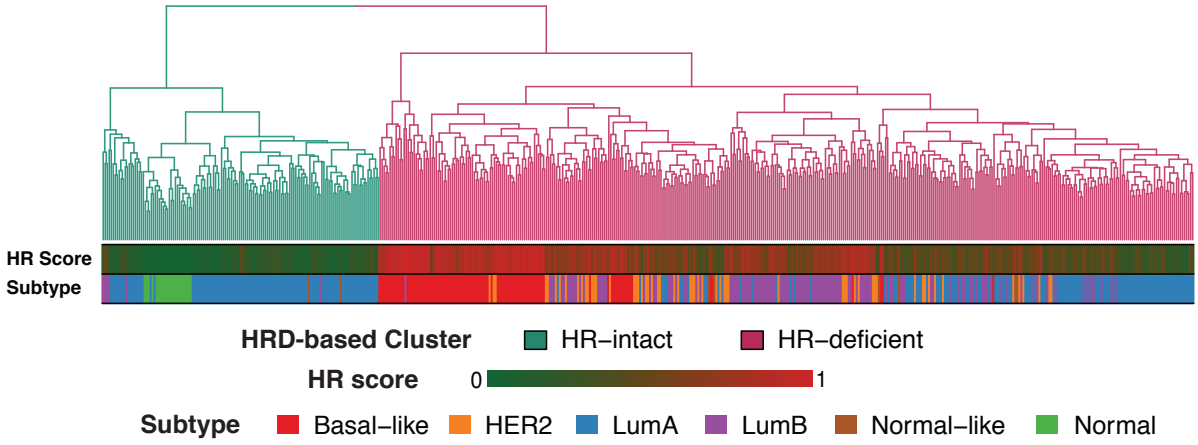


Figure 6 Hierarchical clustering of tumours from the TCGA Microarray cohort based on the HRD signature.

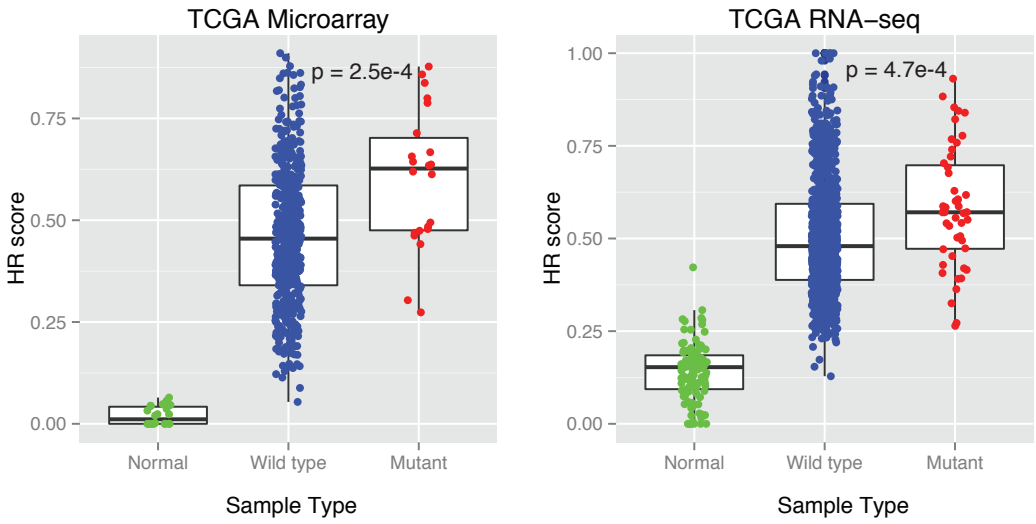


Figure 7 *HR* score versus HR gene mutation for the two TCGA cohorts. *Mutant* refers to tumours with at least one nonsynonymous mutation in any of the six key HR genes (*BRCA1*, *BRCA2*, *RAD51*, *PALB2*, *DNA2* and *EXO1*). *Wild type* refers to tumours with no mutations in these six genes. *Normal* refers to normal breast tissues. P-values were obtained using a Wilcoxon rank-sum test, for the comparison between wild type and mutant tumours.

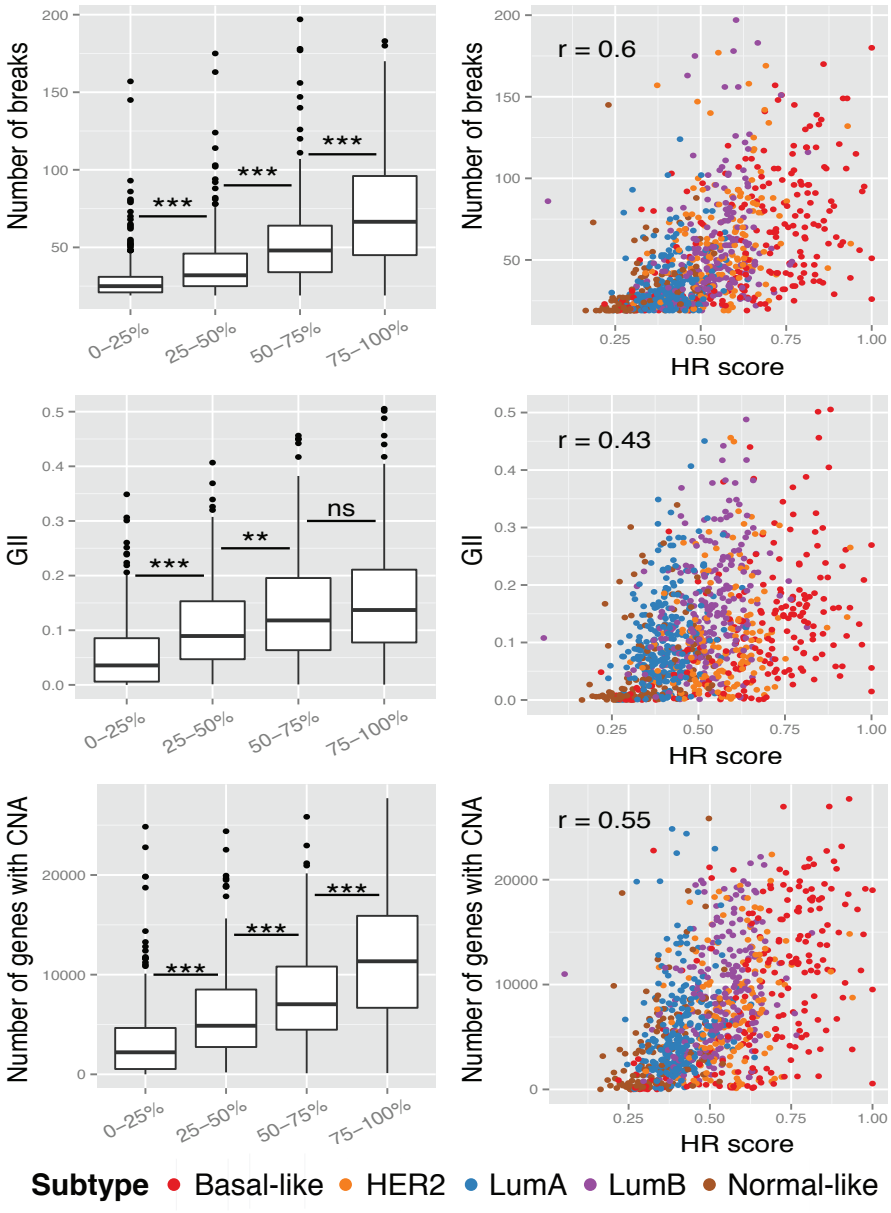


Figure 8 *HR* score versus the three CIN measurements for the METABRIC Validation cohort. Left: Boxplots of the three CIN measurements versus the four *HR* score quartile groups; stars indicate statistical significance according to a Wilcoxon rank-sum test: ns means not significant, ** means 0.001 < p-value <0.01, and *** means p-value < 0.001. Right: Scatter plots of the *HR* score versus each of the three CIN measurements; r represents Pearson Correlation Coefficient.

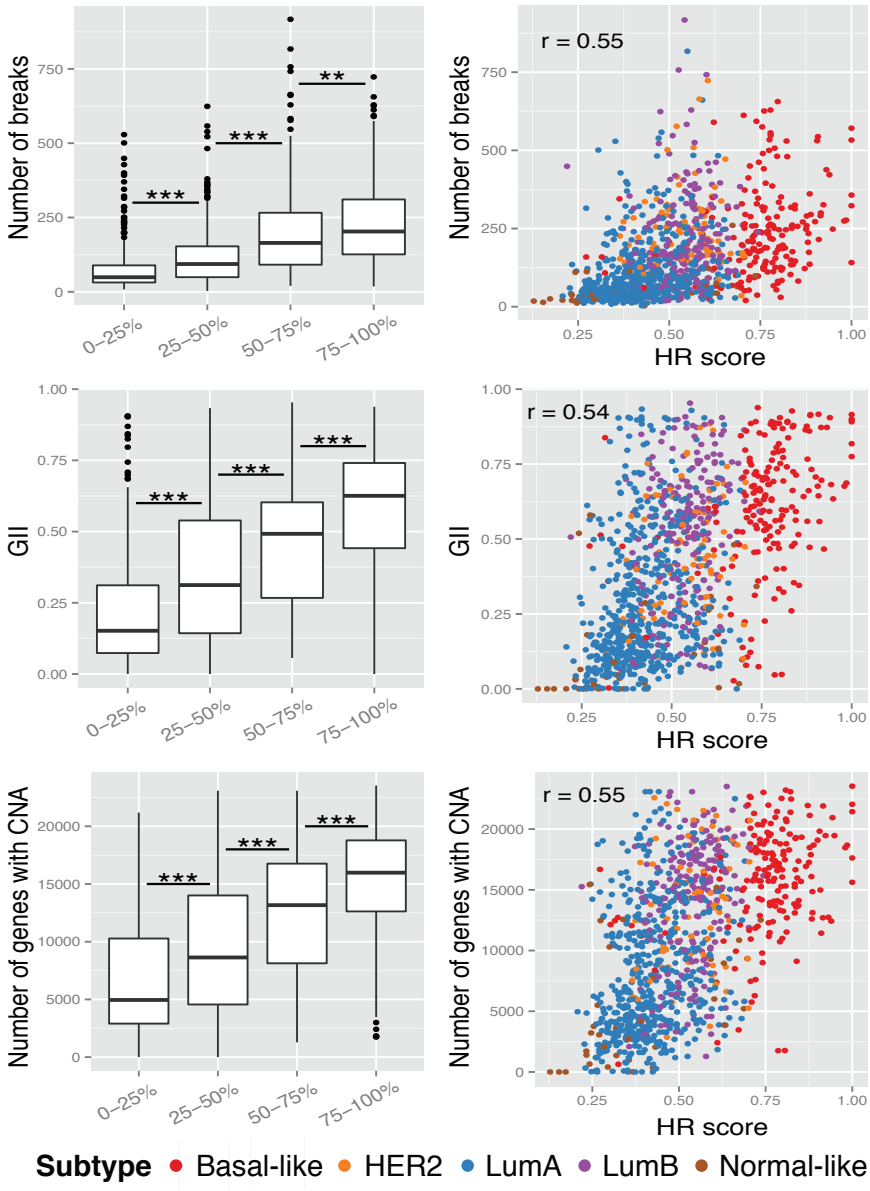


Figure 9 *HR* score versus the three CIN measurements for the TCGA RNA-seq cohort. Left: Boxplots of the three CIN measurements versus the four *HR* score quartile groups; stars indicate statistical significance according to a Wilcoxon rank-sum test: ** means 0.001 < p-value <0.01 and *** means p-value < 0.001. Right: Scatter plots of the *HR* score versus each of the three CIN measurements; r represents Pearson Correlation Coefficient.

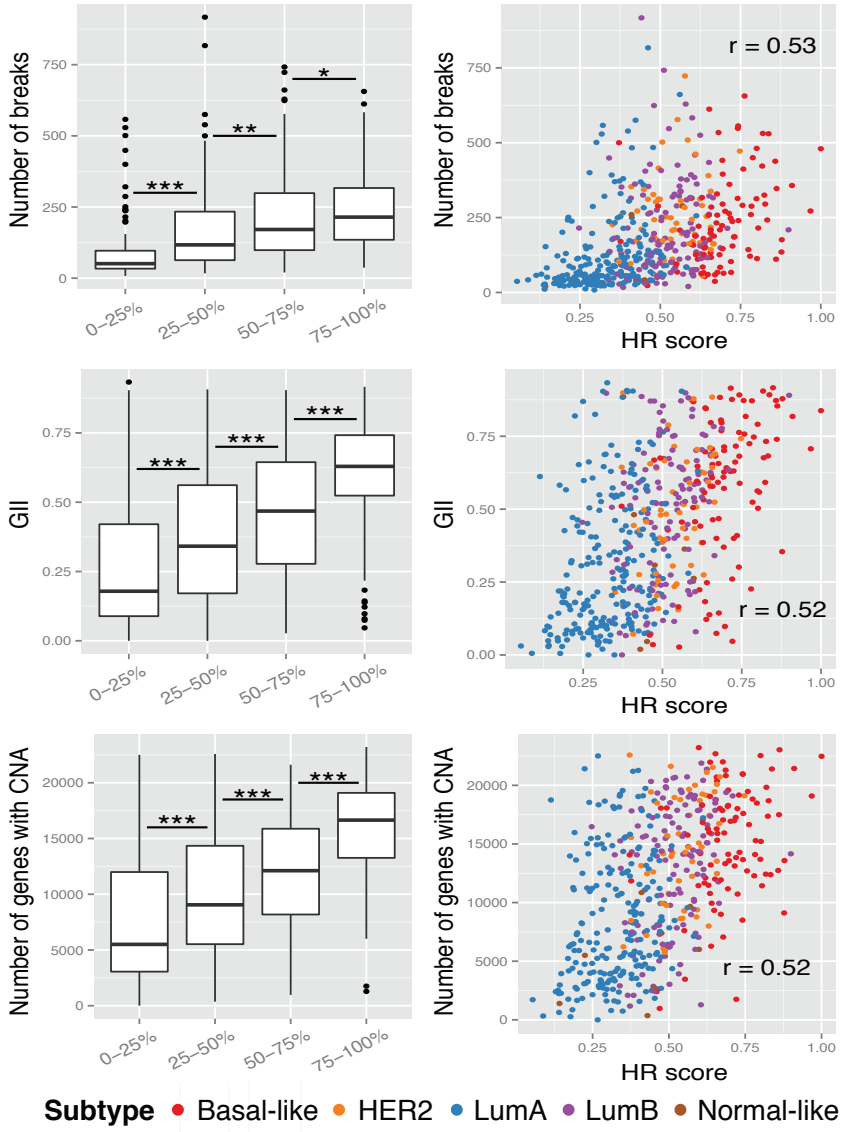


Figure 10 *HR* score versus the three CIN measurements for the TCGA Microarray cohort. Left: Boxplots of the three CIN measurements versus the four *HR* score quartile groups; stars indicate statistical significance according to a Wilcoxon rank-sum test: * means 0.01< p-value < 0.05, ** means 0.001 < p-value <0.01, and *** means p-value < 0.001. Right: Scatter plots of the *HR* score versus each of the three CIN measurements; r represents Pearson Correlation Coefficient.

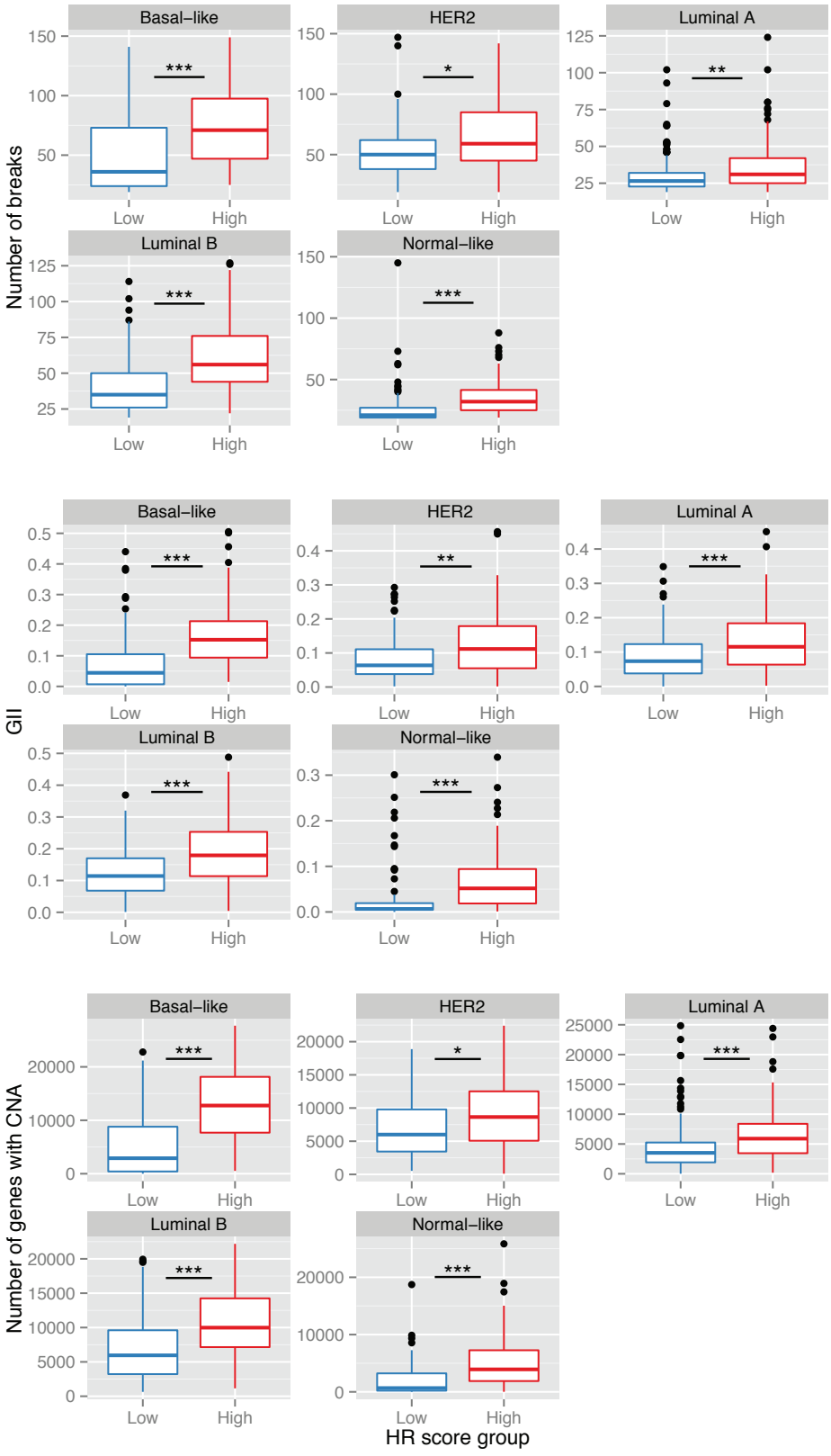


Figure 11 *HR* score versus the three CIN measurements within PAM50 subtypes (METABRIC Validation cohort). For each plot, the two *HR* score groups were divided according to the median *HR* score in each subtype; stars indicate the significance according to a Wilcoxon rank-sum test for each pair of groups: * means $0.01 < p < 0.05$, ** means $0.001 < p < 0.01$, and *** means $p < 0.001$.

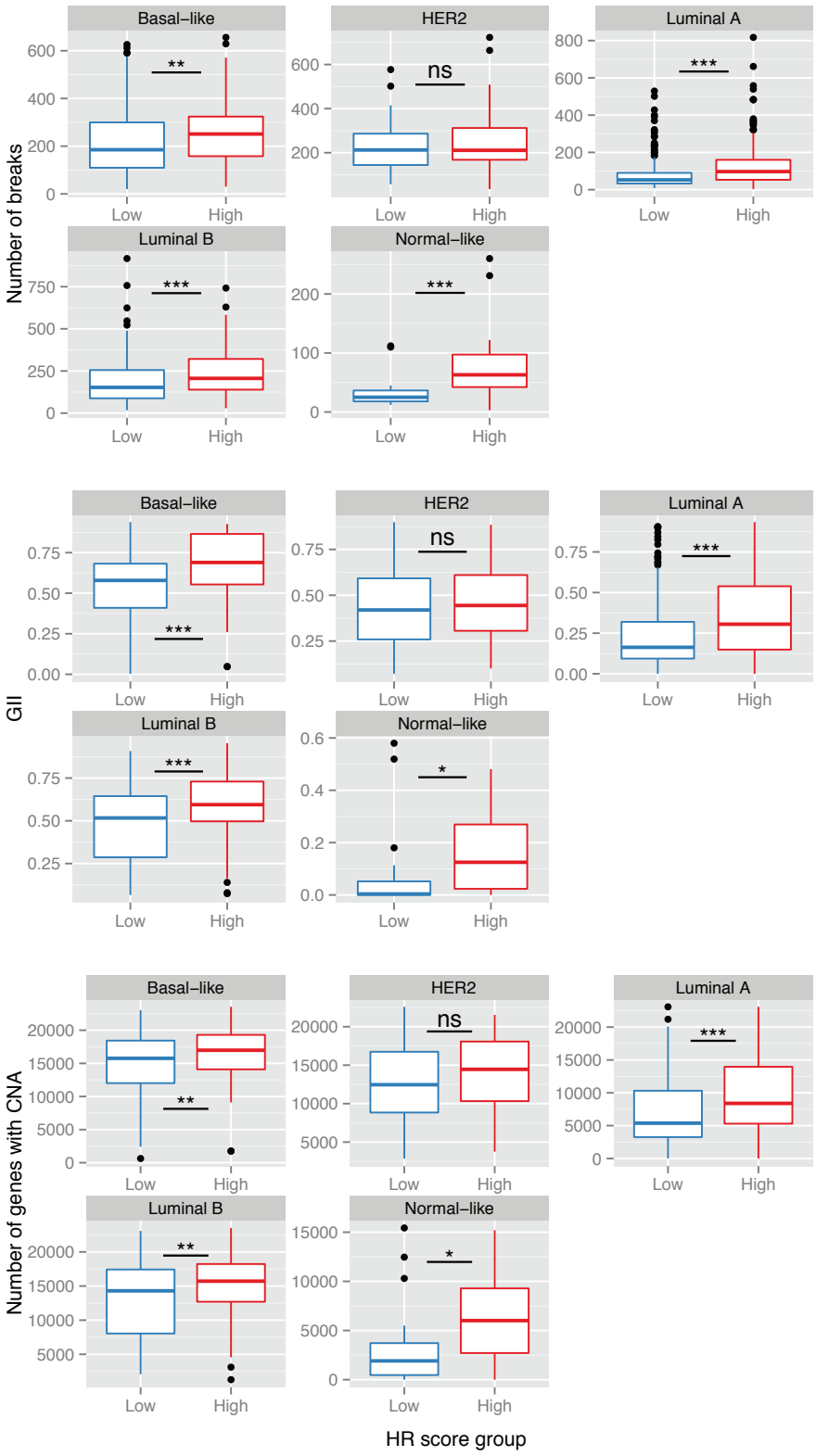


Figure 12 *HR* score versus the three CIN measurements within PAM50 subtypes (TCGA RNA-seq cohort). For each plot, the two *HR* score groups were divided according to the median *HR* score in each subtype; stars indicate the significance according to a Wilcoxon rank-sum test for each pair of groups: ns means not significant, * means 0.01< p < 0.05, ** means 0.001 < p <0.01, and *** means p < 0.001.

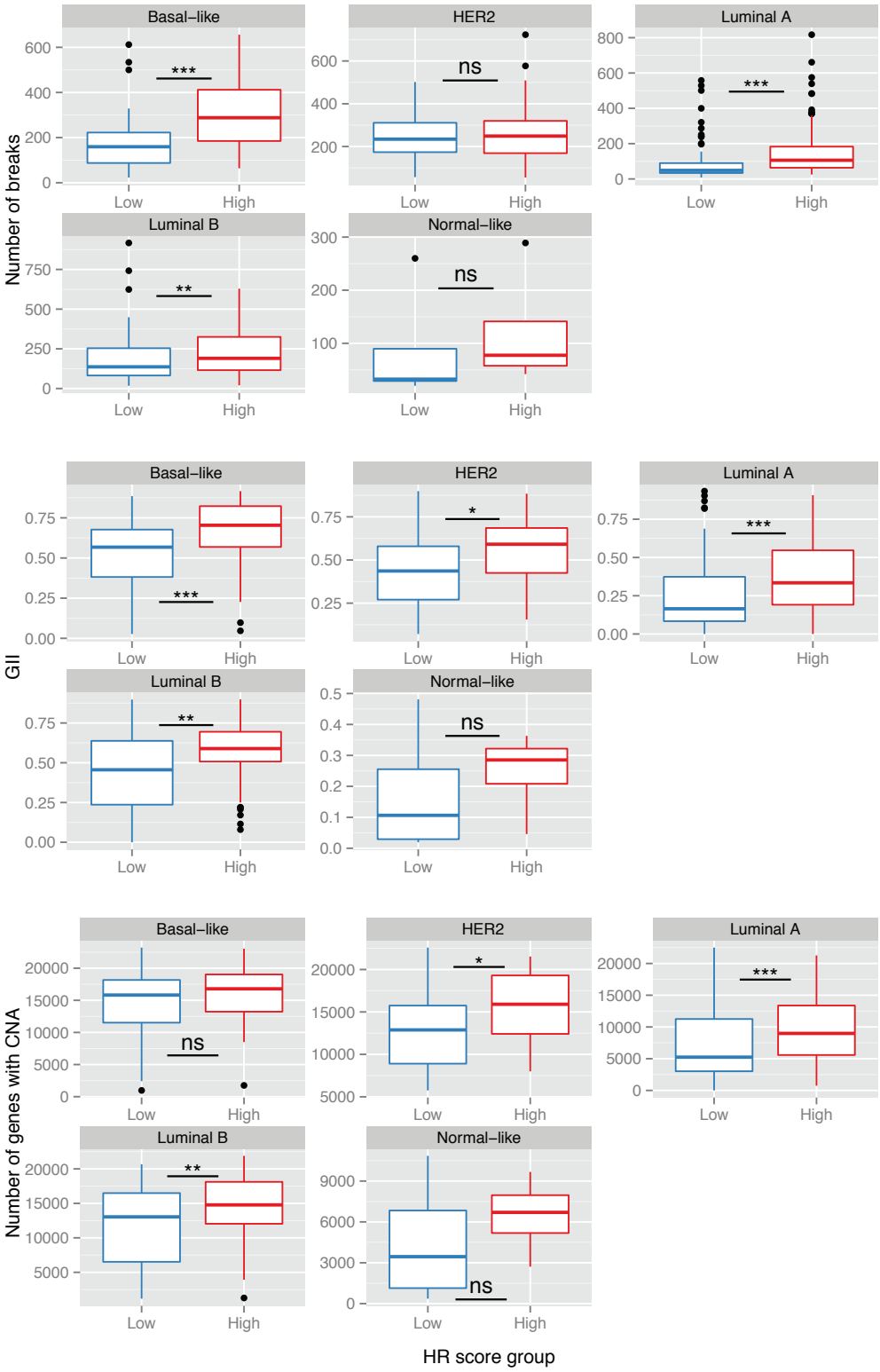


Figure 13 *HR* score versus the three CIN measurements within PAM50 subtypes (TCGA Microarray cohort). For each plot, the two *HR* score groups were divided according to the median *HR* score in each subtype; stars indicate the significance according to a Wilcoxon rank-sum test for each pair of groups: ns means not significant, * means $0.01 < p < 0.05$, ** means $0.001 < p < 0.01$, and *** means $p < 0.001$.

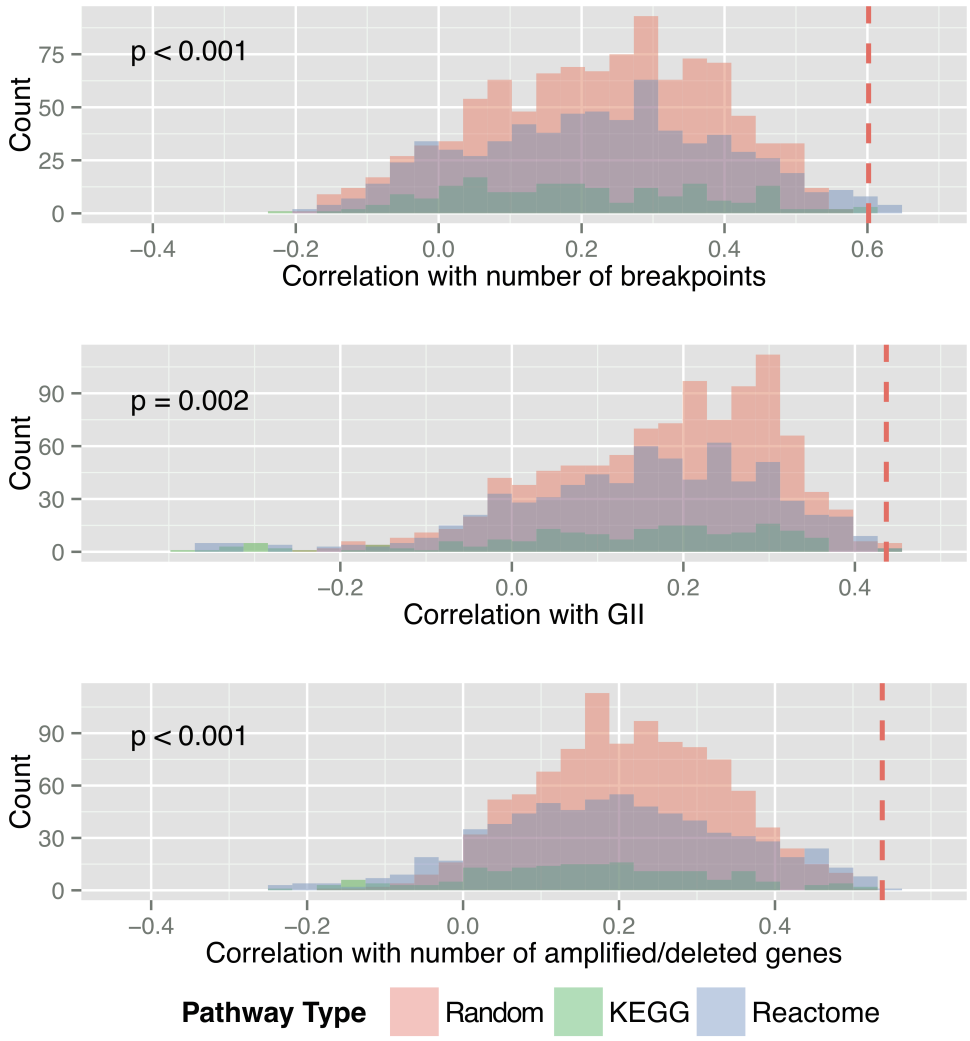


Figure 14 Distributions of the correlations between pathway scores and the three CIN measurements (METABRIC Validation cohort). Results for KEGG pathways are in green, Reactome pathways in blue and Random pathways in pink. Spearman correlation coefficients (r) are represented on the x-axis. Pathway score were calculated with Pathifier. The vertical dashed line in each histogram indicates the value of r between the *HR* score and each of the three CIN measurements, and p represents an empirical p-value for that value of r.

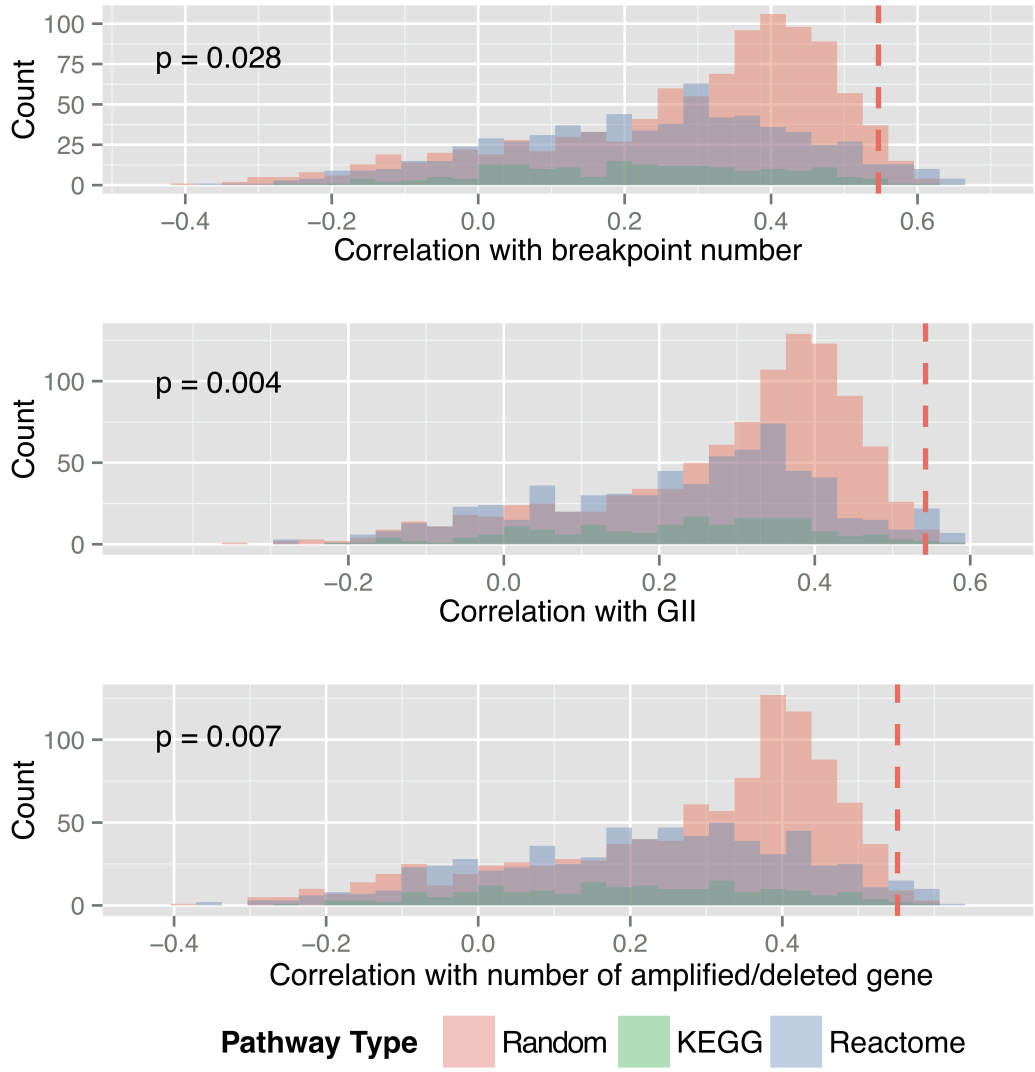


Figure 15 Distributions of the correlations between pathway scores and the three CIN measurements (TCGA RNA-seq cohort). Results for KEGG pathways are in green, Reactome pathways in blue and Random pathways in pink. Spearman correlation coefficients (r) are represented on the x-axis. Pathway score were calculated with Pathifier. The vertical dashed line in each histogram indicates the value of r between the *HR* score and each of the three CIN measurements, and p represents an empirical p-value for that value of r.

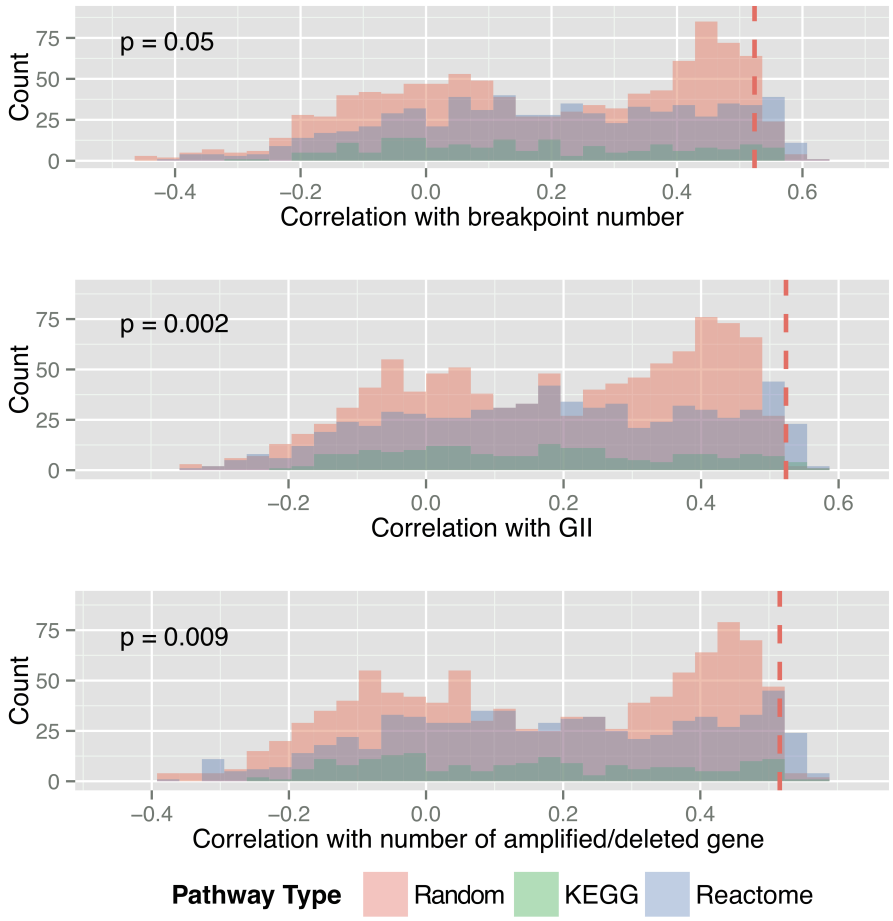


Figure 16 Distributions of the correlations between pathway scores and the three CIN measurements (TCGA Microarray cohort). Results for KEGG pathways are in green, Reactome pathways in blue and Random pathways in pink. An additional 100 CIN-related genes were excluded prior to the construction of the Random pathways as the Pathifer method was sensitive to the addition or removal of a small number of genes in this cohort. Spearman correlation coefficients (r) are represented on the x-axis. Pathway score were calculated with Pathifier. The vertical dashed line in each histogram indicates the value of r between the *HR* score and each of the three CIN measurements, and p represents an empirical p-value for that value of r.