# G-quadruplexes in the Spotlight: Molecular Signatures of Bloom Syndrome

**Dissertation**

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dingwen Su

aus Hubei/China

Tübingen

2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

| | |
|---|---|
| Tag der mündlichen Qualifikation: | 31.07.2024 |
| Dekan: | Prof. Dr. Thilo Stehle |
| 1. Berichterstatter/-in: | Prof. Dr. Yingguang Frank Chan |
| 2. Berichterstatter/-in: | Prof. Dr. Marja Timmermans |

# Acknowledgment

G-quadruplexes stand in one spotlight in my PhD, while the other spotlight shines on the people who have supported me. Without them, this PhD journey could have been considerably more challenging.

First and foremost, I would like to express my gratitude to my supervisor, Frank. Thank you for the supervision, scientific training, and providing the free space to explore things of my own interests! Owing to this, I have gained quite some independence in planning and executing my own project. I also extend my thanks to my TAC members, Detlef and Marja. Thank you for the feedback and inspiration provided during our TAC meetings! Your perspectives have enriched my projects. I would like to sincerely thank Felicity, whom I truly admire as a scientist. Thank you for your insights and bioinformatic tricks shared during our lab meetings!

I owe a great debt to "The Postdok" (or perhaps this time you prefer to be anonymously referred to as Beruška?). You have always been willing to offer professional support whenever I needed it! Reading through my PhD thesis only scratches the surface of it. I am immensely grateful for the tremendous amount of sweets and mandarins. I will never forget your sleepy cake with its secret ingredients (so, what are they?).

Marek, what a coincidence, we have birthday on the same day. Maybe that is why we share preferences on so many things. Thanks for teaching me various things, not limited to experimental skills. You have made my bench work more lively. Moritz and Volker, we started our PhD at roughly the same time and it has been a genuine pleasure. I cannot express enough gratitude for your support through the highs and lows, for your inputs to my project, and for expanding my absolutely random German vocabulary, Kabeltrommel, Alltagssünde, reibungslos... Cholpon, thank you for your continued trust in me and for your encouragement. I love your bulgur salad and the dishes from your country! Julia, your Apfelstrudel is delicious, especially the version with nuts! Thank you for your bioinformatic support as well. Without the aforementioned people, lab life would be mundane and dull! Lisa, your cabbage salad and beetroot salad are exceptional and perhaps the only salads I truly enjoy. Your bean pies are amazing—why can I only have them once a year? Thank you for teaching me how to swim, and I hope I can survive a shark attack! Thanks for pushing me when I'm too shy to try new things or reach out to others. Linda, you've always been incredibly kind, patient, and willing to help. But sorry, you can't put me in your pocket! I am also grateful to other fantastic colleagues and friends at FML: Saudat, James, Dorota and Magda. Thank you all for making FML a lovely community. Without your pranks, humorous

and inspiring minds and the get-together activities, I would have way less fun and laughter!

I want to extend my thanks to Sara. Besides professional support and pranks, thank you for organizing so many food trips!

Special thanks to my bouldering gang: Martina, Lorenzo, Julia (German), Sonja, and Niki! I've had an abundance of fun times with all of you. Martina, you are the brightest person I've ever known! Your positive attitude towards everything has made challenges seem less daunting for me. Thank you for the immersive experience of Italian culture, food, and wine! Lorenzo, thank you for introducing me to bouldering, for the BBQs, and for the incredible pizza! Niki, you make the best soup. The over-night hiking was great, but maybe... a little bit too cold...

Because of COVID-19, I couldn't visit my family often though, I am immensely grateful to them, especially my brother, my mother, and my grandma. From the thesis guidelines, I couldn't find any restrictions stating that I cannot write acknowledgments in Chinese. My lovely mother can barely read English, although she has declared a thousand times that she will start learning it. To save her the effort of translating the relevant paragraphs into Chinese, I will write them in Chinese. 感谢龚师傅一路以来的

支持与鼓励。龚师傅践行着“路漫漫其修远兮，吾将上下而求索”。她对于生活的勇气，对新事物的开放和乐于探索的态度，终身学习的心态以及对她热爱的事物的热诚是我的榜样，很幸运我们也是亲近可以平和地讨论生活和交换意见的朋友。

I am particularly grateful to my dear friends: Wenxuan Guo (郭文煊), Yinan Wang (王祎楠), Dr. Yingjing Miao (苗英靖), Sixiu Zhao (赵思修), and Hangning Weng (翁航宁).

Thank you for encouraging me, for taking care of me, for listening to my complaints, for sharing my joy, and for feeding me! You guys created a huge comfort zone for me. It is a true blessing to have known you all!

Sincere thanks extend to Christopher M. Cunniff and Maeve Flanagan at the Bloom Syndrome Registry at Cornell Weill Medicine. Importantly, we are all deeply indebted to the Bloom Syndrome families for their contributions to medicine and science. I am also grateful for the support from the institute, genome center, and international office, especially George and Susan for their help with my visa application. Additionally, I thank Rebecca for arranging an office for us!

As it for the PhD, finally, *per aspera ad astra*.
As it for the past and for the future, *Homo, nosce te ipsum*.

# Table of contents

# Summary

Bloom Syndrome (BS) is a recessive genetic disorder characterized by hyper-recombination and genome instability. It is caused by mutations in the conserved RecQ helicase gene, BLM, which is essential in maintaining genome integrity given its ability to unwind various aberrant DNA structures arising in DNA metabolic pathways, including a four-stranded helical structure called G-quadruplexes (G4). Clinically, BS individuals manifest developmental delay, immune deficiencies, a shorter lifespan and elevated cancer risks. Molecularly, BS is characterized by excessive sister chromatid exchange (SCE) events and genome instability. Recent studies hinted at the relevance of G4 in BS by the enrichment of putative G4-forming sequences in the differentially expressed genes and SCE events. However, having G4-forming sequences does not predict the formation of G4 in the cells, and it remains unclear if endogenous G4 structures play a pivotal role in BS. Given G4's regulatory roles in DNA replication, transcription, and chromatin organization, I hypothesized that BLM deficiency leads to perturbed G4 formation and resolution, causing the downstream molecular changes in BS.

To characterize the molecular changes in BS, we collected cell lines derived from sex-matched and roughly age-matched BS and healthy donors (WT). I assayed the molecular changes via ATAC-seq and RNA-seq. Importantly, instead of inferring the G4 formation by *in silico* predicted G4 motifs or *in vitro* validated G4-seq hits, I utilized G4 ChIP-seq and an antibody specific to G4 structures to map the endogenous G4. I found that in BS, increased G4 formation correlated with both increased chromatin accessibility and gene expression, and vice versa. Next, by applying a G4 stabilizing molecule to WT to mimic the defective G4-resolving abilities in BS, I showed that G4 stabilization partially phenocopies BS across both cell types, suggesting a partially causal role of G4. To dissect the molecular mechanism, I observed that regions/genes with the presence of G4 displayed elevated chromatin accessibility and gene expression. My results suggest a plausible molecular mechanism wherein changes in G4 formation directly influence local chromatin accessibility, thereby regulating gene expression. As a validation for this hypothesis, I collected ATAC-seq data from a BS

family and showed that more accessible chromatin regions in BS individuals exhibited stronger enrichment for G4-forming sequences.

During data analysis, I noticed the impacts of copy number variation (CNV) in differential analyses. In chapter two, I demonstrated that CNV between contrasted samples can and does drive much of the observed differential signals. However, the impact of CNV is often overlooked despite it being common in diseases like cancer. I therefore developed a differential analysis pipeline featuring copy number (CN) normalization. I showcased its application and advantages using two examples in biomedical studies. Firstly, I applied it to ATAC-seq and ChIP-seq data generated from cell lines with complex chromosomal aberrations derived from a BS individual and a healthy donor. Using a conventional copy-number blind pipeline, differential signals were heavily skewed toward the sample with relatively higher copy numbers in the corresponding regions. Notably, applying our pipeline with CN normalization efficiently distinguishes differential signals driven by CNV and those due to the disease. In the second case, I applied it to ATAC-seq data generated from trisomy 21 and euploid cell lines. By combining the results from our pipeline with the common workflow, I was able to distinguish among open chromatin regions on chromosome 21 with dosage effects, compensatory effects and copy-number-independent regulatory changes.

Overall, this thesis presents the first study with direct evidence that G4 structures could emerge as a central factor in the molecular etiology of BS. Upon the loss of function of BLM, defective G4-resolving abilities lead to G4 formation changes, which subsequently trigger downstream molecular alterations in BS, thereby contributing to the clinical phenotypes observed in affected individuals. My findings enrich the understanding of Bloom Syndrome at a molecular level as well as the regulatory function of G4 *in vivo* and highlight the broad regulatory function of BLM, which extends beyond its well-established role as a helicase. Additionally, this thesis delivers the concept of CN normalization for differential analyses of count-based functional genomic assays.

# Zusammenfassung

Bloom-Syndrom (BS) ist eine rezessive genetische Erkrankung, die sich durch Hyper-Rekombination und genomische Instabilität auszeichnet. Die Erkrankung wird durch Mutationen im konservierten RecQ-Helikase-Gen *BLM* verursacht, welches essentiell für die Aufrechterhaltung der genomischen Integrität ist, da es abnormale DNA-Strukturen in Stoffwechselwegen auflöst, einschließlich viersträngiger helikaler Strukturen names G-Quadruplex (G4). Klinisch manifestiert sich BS in Patienten mit Entwicklungsverzögerungen, Immundefizienzen, einer verkürzten Lebensdauer und einem erhöhten Krebsrisiko. Molekular ist BS durch einen übermäßig häufigen Schwesterchromatidaustausch (SCE) und genomische Instabilität gekennzeichnet. Neueste Studien deuteten auf eine Relevanz von G4 im Bezug auf BS hin, da potenziell G4-bildende Sequenzen in den differentiell exprimierten Genen und SCE-Ereignissen angereichert sind. Allerdings lässt sich von G4-bildenden Sequenzen nicht zwangsläufig die Bildung von G4 in den Zellen schließen und es ist unklar, ob endogene G4-Strukturen eine entscheidende Rolle im BS spielen. Angesichts der regulatorischen Funktionen von G4 in DNA-Replikation, Transkription und Chromatinorganisation, habe ich die Hypothese aufgestellt, dass eine BLM-Defizienz zu gestörten G4-Formationen und -Auflösungen führt, was die nachfolgenden molekularen Veränderungen des BS verursacht.

Zur Charakterisierung der molekularen Veränderungen des BS benutzten wir zwei Paare von Zelllinien, die von gleichgeschlechtlichen und ungefähr gleichaltrigen BS- und gesunden Spendern (WT) stammten. Ich analysierte die molekularen Veränderungen mittels ATAC-Seq und RNA-seq. Anstatt jedoch auf G4 Präsenz durch *in silico* vorhergesagte G4 Motive oder *in vitro* validierte G4-Seq Treffer zu schließen, habe ich G4 ChIP-seq zusammen mit einem G4-spezifischen Antikörper benutzt um endogene G4 Strukturen zu kartieren. In BS-Proben korrelierte die differentielle G4-Formation positiv mit erhöhter Chromatin-Zugänglichkeit und Genexpression und umgekehrt. Durch die Verwendung eines G4-stabilisierenden Moleküls in WT Zellen, welches die defekte G4 Auflösung im BS imitiert zeige ich, dass die Stabilisierung von G4 ungeachtet des Zelltypes teilweise BS phänotypisch nachahmt, was auf eine

kausale Rolle von G4 im BS hindeutet. Beispielsweise zeigten Regionen/Gene in der Anwesenheit von G4 eine höhere Chromatin-Zugänglichkeit und Genexpression. Möglicherweise beeinflussen Veränderungen in der G4-Bildung in BS direkt die lokale Chromatin-Zugänglichkeit und regulieren so die Genexpression. Um dieser Hypothese zu bestätigen, generierte ich ATAC-Seq Daten von einer BS-Familie und zeigte, dass in BS-Patienten zugänglichere Chromatinregionen eine stärkere Anreicherung von G4-bildenden Sequenzen aufwiesen.

Während der Datenanalyse bemerkte ich, wie Copy Number Variation (CNV) die differentiellen Analysen beeinflusste. In Kapitel Zwei habe ich gezeigt, dass CNV einen Großteil der beobachteten differentiellen Signale zwischen Proben antreiben kann, der Einfluss von CNV auf die Analysen allerdings oft nicht beachtet wird, obwohl er bei Krankheiten wie Krebs häufig vorkommt. Deshalb habe ich eine Pipeline entwickelt, die Copy Number (CN) Normalisierung während differentieller Analysen umfasst. Ich zeigte ihre Anwendung und Vorteile anhand von zwei Beispielen in biomedizinischen Studien auf. Zunächst wandte ich sie auf ATAC-seq und ChIP-seq Daten an, die von Zelllinien aus einem BS-Patienten sowie einem gesunden Spender mit komplexen chromosomalen Aberrationen generiert wurden. Mit einer herkömmlichen Pipeline waren differentielle Signale stark auf die Probe mit relativ höheren Kopienzahlen in den entsprechenden Regionen verzerrt. Die Anwendung unserer Pipeline mit CN-Normalisierung unterschied effizient zwischen differentiellen Signalen, die durch CNV verursacht werden, und solchen, die tatsächlich auf die Krankheit zurückzuführen sind. Im zweiten Fall habe ich die Pipeline auf ATAC-Seq-Daten angewandt, die aus Trisomie 21 und euploiden Zelllinien generiert wurden. Durch die Kombination unserer Pipeline mit dem herkömmlichen Arbeitsablauf konnte ich zwischen den Chromatinregionen auf Chromosom 21 mit Dosierungseffekten, kompensatorischen Effekten und regulatorischen Veränderungen unabhängig von der Anzahl der Kopien unterscheiden.

Insgesamt repräsentiert diese Thesis die erste Studie mit direkten Beweisen dafür, dass G4-Strukturen als zentraler Faktor in der molekularen Ätiologie von BS auftreten können. Durch den Funktionsverlust von BLM führen defekte G4-auflösende Fähigkeiten zu Veränderungen in der G4-Bildung, die anschließend nachgelagerte molekulare Veränderungen in BS auslösen und damit zu den klinischen Phänotypen

beitragen, die bei betroffenen Personen beobachtet werden. Meine Ergebnisse erweitern das Verständnis des Bloom-Syndroms auf molekularer Ebene sowie die regulatorische Funktion von G4 *in vivo* und heben die weitreichende regulatorische Funktion von BLM hervor, die über seine etablierte Rolle als Helikase hinausgeht. Darüber hinaus liefert diese Arbeit das Konzept der CN-Normalisierung für differentielle Analysen von zahlenbasierten funktionellen genomischen Assays.

# Introduction

## 1. Motivation and context

### 1.1. G4-forming sequences and Bloom Syndrome

Bloom Syndrome (BS), is a rare autosomal recessive disorder caused by loss-of-function mutations in the *BLM* gene [1]. It was first described in 1954 by Dr. David Bloom and *BLM* was identified as the underlying risk gene 40 years later [1,2]. BLM helicase is a conserved helicase that exists in *E. coli*, yeasts, drosophila, mice, and humans, as well as in plants [3]. It processes and unwinds various aberrant DNA structures (e.g., DNA G-quadruplexes) and thus plays an essential role in maintaining genome integrity [3–6]. Particularly, it is indispensable in homologous recombination (HR) dependent DNA repair and in suppressing illegitimate DNA recombination events [5,7,8]. The dysfunction of the BLM has detrimental impacts on the affected individual and cells. BS individuals have complex clinical manifestations. They displayed prenatal and postnatal growth deficiency, immune deficiency and strikingly reduced lifespan[9–11]. Likely due to the lack of BLM to maintain genome stability, BS patients are prone to develop multiple cancers at an early age [9,12]. Emerging studies have been carried out to understand the molecular changes in BS and gain insight into the etiology of BS. At the molecular level, BS cells show characteristically excessive sister chromatin exchange (SCE) events [13]. They display hallmarks of genome instability and elevated frequencies of loss-of-heterozygosity, which increases the risk of developing cancer [11,14–16]. Additionally, BS cells display a slower cell cycle and altered gene expression profiles [17,18].

G-quadruplexes (G4s), one of BLM's substrates, are four-stranded non-canonical DNA structures formed by G-rich sequences. Bioinformatic tools predicted over 400,000 up to one million G4 motifs in the human genome [19]. Another approach, G4-seq validated more than 500,000 G4-forming sequences in the purified human genomic DNA [20]. Recent advances in screening G4-specific antibodies and genomic techniques such as G4 ChIP-seq have enabled mapping endogenous G4s in living cells and it identifies about 10,000 to 20,000 G4s [21–24]. They reside mostly in open

chromatin regions and promotes, and have broad regulatory functions in DNA replication, transcription, and DNA-protein interactions [21–24].

Recent studies hinted at the relevance of G4 in BS. Van Wietmarschen et al. mapped the SCEs at the single-cell level and showed that SCEs in wild-type human and murine cells did not tend to have a hotspot whereas in BS cells they were enriched for G4 motifs, particularly at the transcribed genes [25]. Moreover, differentially expressed genes in BS were correlated with the presence of G4 motifs [18,26]. However, after expanding the sample size, another study argued that there was no enrichment of G4-seq hits in the differentially expressed genes in BS [27].

### 1.2. Aim of study

These possible connections notwithstanding, the described association between molecular changes and G4 in BS is limited to putative G4-forming sequences. Crucially, neither *in silico* predicted G4 motifs nor *in vitro* validated G4-seq reliably predict the G4-forming status *in vivo*. If the local presence of G4 changes in BS and whether these changes are connected to the molecular changes in BS remains unknown. In addition, RNA-array was used in previous studies to profile the transcriptome. This study aims to characterize the molecular changes via RNA-seq and ATAC-seq, map endogenous G4 in wild-type and BS cells and address the significance of G4 in BS.

## 2. Human RecQ helicases

Helicases are the ubiquitously expressed enzymes that unwind double-stranded DNA (dsDNA) into single-stranded DNA (ssDNA) and/or unfold secondary structures formed within a DNA strand or RNA strand during various fundamental and essential cellular processes, such as DNA replication, transcription, DNA repair and maybe even in translation. RecQ helicases, named after the recQ gene of *E. coli*, are proteins conserved from bacteria to humans and play crucial roles in genome maintenance and integrity for which they are often referred to as the guardians of the genome. Most mammals have five RecQ helicases: ATP-dependent DNA helicase Q1 (RECQ1), RECQ4, RECQ5, BLM, and WRN (**Fig. 1a**). These RecQ helicases share conserved domains and work cooperatively in multiple processes to maintain genome integrity

(**Fig. 1**). In certain DNA metabolic pathways, they complement each other and each member has unique sets of protein-interacting partners dictating their specialized functions in different DNA metabolic pathways (**Fig. 1b**) [4,5,28]. They can play both overlapping and non-redundant roles. Defects in human RECQ4, BLM, and WRN cause rare and severe autosomal recessive cancer predisposition disorders – Rothmund-Thomson, Bloom, and Werner syndrome respectively. All RecQ helicases are ATP-dependent 3' - 5' DNA helicases and can unwind a variety of DNA substrates (**Fig. 2**) [5,29]. The 3' - 5' directionality indicates that RecQ helicases prefer substrate with a 3' single-stranded (ss) extension which they are thought to bind to and translocate along. Nevertheless, different RecQ helicases can untangle various aberrant DNA structures and show preferences towards specific substrates [5,30].

# 3. BLM helicase

## 3.1. The DNA substrates of BLM helicase

Bloom Syndrome RecQ-like helicase *BLM* gene is mapped to chromosome 15q26.1 and the only known causal gene underlying BS (OMIM 210900). It encodes BLM protein which belongs to the RecQ helicase subfamily [1,29]. BLM protein has no helicase activity on double-stranded DNA but it displays a strong helicase activity when acting on more elaborate DNA structures such as bubble structures caused by DNA mismatch, displacement loop (D-loop), G-quadruplex (G4), or Holliday junction (HJ) [4,6,30,30–32]. D-loop is a structure where two strands of dsDNA are focally separated due to the invasion of a third ssDNA which is homologous to and pairs with one of the strands. D-loops naturally form at telomeres to stabilize them. They also exist as intermediates in HR and DNA replication. If not disrupted, the extension of a D-loop eventually generates HJs. Noticeably, among all the RecQ helicases WRN and BLM both can resolve HJs, but only BLM can dissolve double HJs, implicating its unique and important roles in recombination. G-quadruplexes are non-Watson-Crick highly stable four-stranded helical DNA, DNA-RNA hybrid structures or RNA formed by guanine-rich sequences. Sequences with the potential to form G4 structures are abundant in G-rich telomere regions in human. G4 structures may interfere with DNA replication, transcription and translation; thus, they are thought to have regulatory roles

in gene expression [33,34].

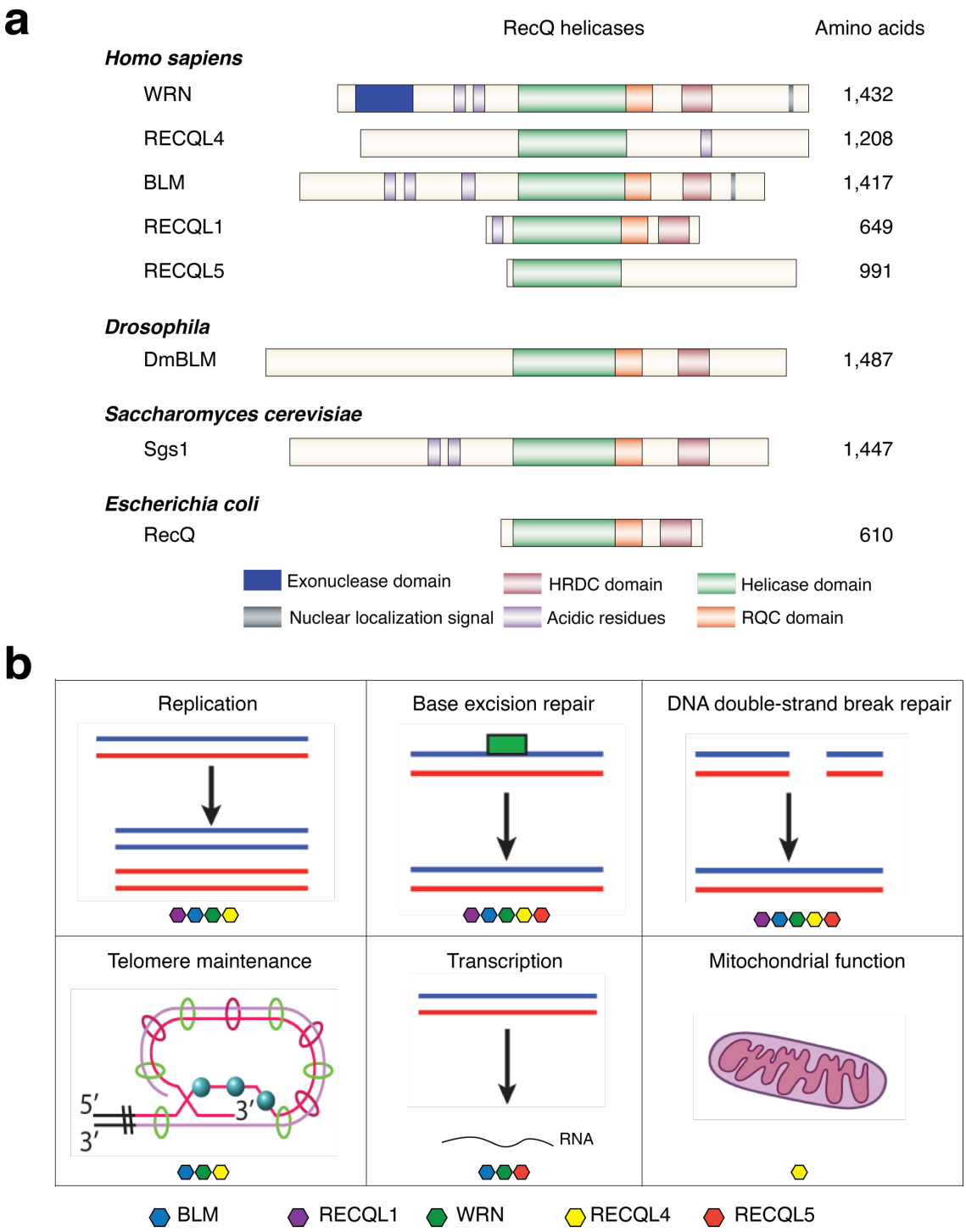



**Figure 1. The structures and functions of human RecQ helicases.** (**a**) Schematic representation of the structures of human RecQ helicases. (**b**) Functions of human RecQ helicases. Each RecQ helicase is represented as a colored hexagon and hexagons at the bottom of each process indicate the protein involved in the process. BER, base excision repair; DSBR, double-strand break repair; mRNA, messenger RNA. Figures are adapted from [5,28].

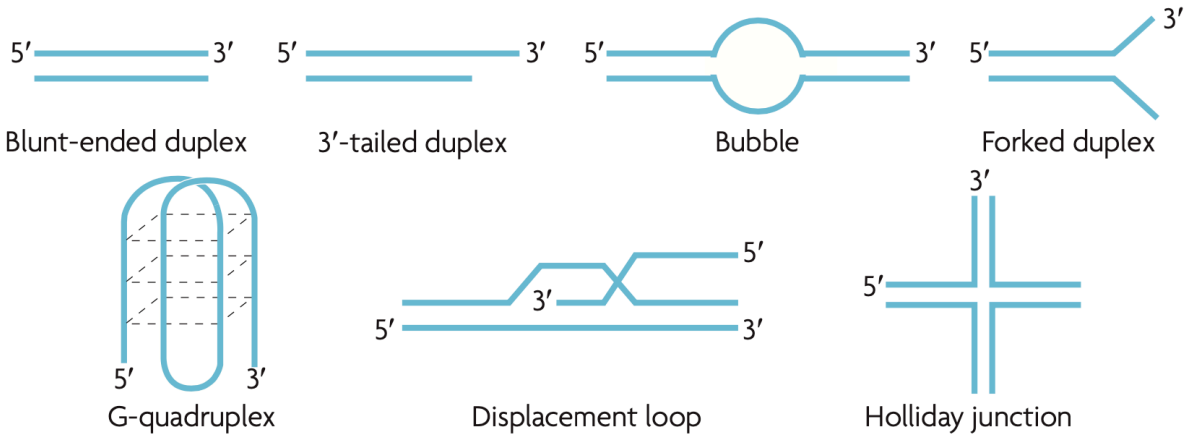



**Figure 2. Substrates of RecQ helicases.** Bubble structures arise from DNA mismatch. Forked duplexes occur during DNA replication. 3'-tailed duplex (3' single-strand DNA overhang), displacement loop and Holliday junction are intermediates in the homologous recombination process. The figure is taken from [4].

## 3.2. The structure of BLM helicase

BLM protein contains 1417 amino acids and consists of a well-conserved helicase core and ancillary N- and C-terminal domains flanking the helicase core. It has been postulated that significant differences in the N-terminal domains and C-terminal extensions might explain functional differences among family members [35]. However, these domains remain poorly characterized. Among the other four RecQ helicases, BLM shares the most structural and functional similarities with WRN [5,35,36].

The N-terminal domain of BLM is found to be involved in the dimerization and oligomerization of BLM. Regardless of the N-terminal domain of RecQ helicases lacking the general sequence and structural conservation, a conserved protein fragment termed as the dimerization helical bundle in the N-terminal domain (DHBN), was characterized at their N-terminus. This domain, proposed to have a 2–3 alpha-helix conformation, mediates dimer and potentially hexamer formation [37]. The N-terminus of BLM also serves as a target for some post-translational modifications such as phosphorylation and SUMOylation, which can regulate its interaction with other proteins [38]. The helicase core comprises the most important functional domain of the helicase. It consists of three highly conserved and important functional domains: the helicase domain, the RecQ carboxyl-terminal (RQC) domain, and the Helicase and RNaseD C-terminal domain (HRDC) [36,39]. This helicase domain defines the RecQ

family and participates in coupling the energy of NTP hydrolysis to unwinding nucleic acid duplexes [40]. It contains seven conserved helicase motifs (I, Ia, II, III, IV, V, VI) that are commonly seen in DNA and RNA helicases [41]. Helicase motifs I and II serve as sites interacting with ATP and magnesium ions. Besides these classical helicase motifs, BLM possesses an element called 'motif 0', which is N-terminal to motif I and crucial for the helicase function [41,42]. A missense mutation in motif 0 was found to cause Bloom Syndrome in human and the corresponding murine genetic variant abolishes the ATPase and helicase activity [43,44].

The RQC domain is unique to RecQ family and it specifically mediates protein binding to G4 structures and stabilizes the binding to other DNA substrates [45]. It consists of a zinc-binding motif with four conserved cysteine residues, a helix-hairpin-helix, a winged-helix (WH) subdomain, and a β-hairpin motif revealed by X-ray crystallography [5]. The purified WH subdomain of BLM and WRN displays strong DNA binding activity *in vitro* [36,46]. Interestingly, although HDRC domain exists in *E. coli* RecQ and *S. cerevisiae* Sgs1, only two human RecQ helicases, BLM and WRN, possess this domain, suggesting that it may mediate special functions of BLM and WRN in human.

The BLM HRDC domain consists of five α-helices with an extended acidic surface formed by residues in helices 3–5. This domain is of great importance to telomere maintenance and suppresses illegitimate recombination events during DNA replication. It was proposed that this domain may act to separate the junction sites of dHJ structures via electric repulsion between the acidic surfaces when BLM oligomerizes [47]. Interestingly, both the RQC domain and HRDC domain interact with the telomere-associated protein TRF2 in Shelterin complex, which stabilizes and protects the ends of telomeres during DNA replication [41].

## 3.3. The expression and localization of BLM helicase

BLM is abundant in rapidly dividing cells such as cancer cells and is not expressed in non-dividing cells. The expression of BLM is cell cycle regulated. It accumulates in S phase, persists in G2/M phases and sharply declines to an undetectable level in G1 phase through a ubiquitin-proteasome independent pathway, strongly suggesting a role in DNA replication and HR which occurs only in the S and G2 phases [17]. Among different tissues, *BLM* mRNA is highly expressed in lymphoid tissues (thymus, bone

marrow, spleen, lymph node) and salivary gland [48]. The protein is detected in high abundance in most of the organs/tissues, such as lymphoid tissues, brain, respiratory system, muscle, skin, testis and ovary (The human protein atlas) [48].

Regarding cellular localization, BLM is a nuclear protein localized in promyelocytic leukemia bodies and telomeres [49,50]. The nuclear localization signal is located in its C-terminus [51,52]. Protein truncation mutations in BLM leaving out the C-terminus result in the inability of mutant BLM to translocate into the nucleus [52].

## 3.4. The function of BLM helicase in DNA replication

Many studies have demonstrated the substantial role of BLM in maintaining genome integrity, putatively acting as the caretaker of the genome. Involved in crucial steps of HR, BLM is, therefore, a key player in various processes in DNA replication and DNA repair [4,5,28,41].

The process of DNA replication is stressful and fraught with potential dangers. DNA lesions, proteins bound to the template strand, secondary structures, and topological stress can all inhibit replisome movement leading to replication fork stalling or even collapse. BLM is capable of tackling these potential events and therefore facilitates the progression of replication (**Fig. 3**) [5,28,53–59]. It is able to untangle aberrant DNA secondary structures that impede fork progression and resolve blocked forks when there are obstacles on the DNA strands. HR is a commonly used mechanism to restart stalled or collapsed replication forks. In the case of encountering a blocking lesion in the leading-strand template, the replication fork may regress by WRN and/or BLM. If the synthesis of lagging strands continues beyond the point where the fork was arrested, nascent DNA strands anneal forming a four-way junction that is structurally analogous to HJs (**Fig. 3c**). BLM can help bypass the lesion and reset the fork by facilitating reverse branch migration. Alternatively, the four-way junction might be cleaved leaving broken-ended DNA and the collapsed replication fork. A similar intermediate might arise when a fork meets ssDNA discontinuity (**Fig. 3d**). In these scenarios, replication can be reinitiated using HR-dependent mechanism – break-induced replication (BIR). The processes of ssDNA end-resection, RAD51 filament formation and homology searching are the same in BIR and HR. The critical difference between BIR and the HR-dependent dHJ mechanism is that there is only one DNA

strand invasion and both leading strand and lagging strand synthesis in BIR whereas dHJ mechanism involves the invasion of both strands and both nascent strands serve as primers setting up leading strand synthesis. Occasionally, DSBs might be generated if two forks progress to a nick or a gap in between, the repair of which will be elaborated later (**Fig. 3e**).

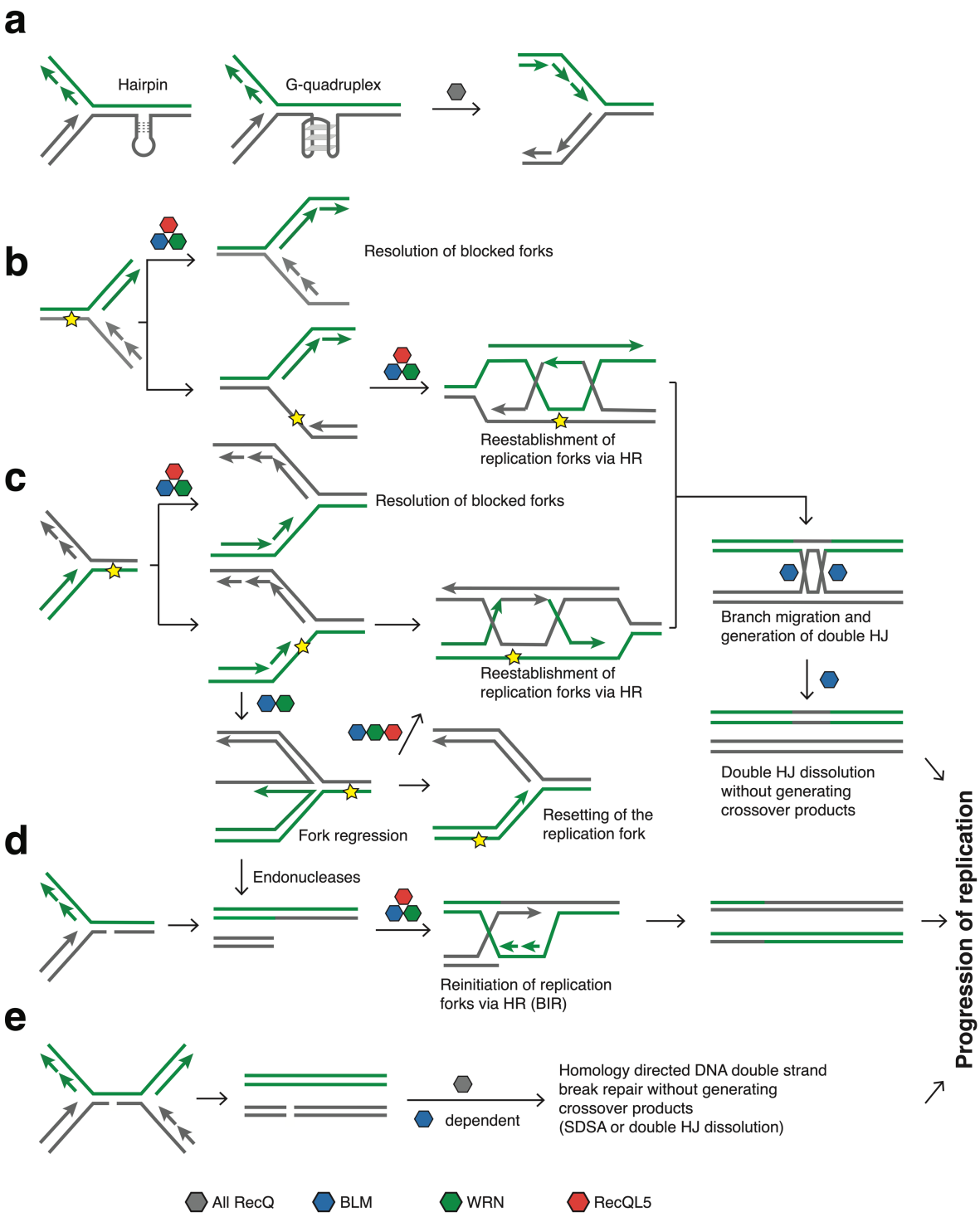



**Figure 3. Functions of BLM in maintaining genome integrity during replication.** (**a**) BLM

and other RecQ helicases unwind aberrant DNA structures. (**b**) When there is a barrier or lesion on the leading strand, BLM is able to remove it (upper panel) or reestablish the stalled replication forks by HR forming a D-loop (bottom panel). (**c**) When there is a barrier or lesion on the lagging strand, BLM resolves it and the replication continues. If the replication fork is stalled, BLM is involved in reestablishing the replication fork by HR or resetting the replication fork leading to the bypass of the lesion by migrating the four-way junction after fork regression. The regressed fork might also be cleaved by endonucleases and completely disrupted. (**d**) Nicks or gaps on one parental strand of the replication fork cause the collapse of the replication fork. Broken forks might be repaired by BIR. (The representation here only shows the cases of discontinuity in the leading strand). (**e**) Two replication forks meet the nick or a gap in the middle resulting in a DSB, which might be repaired by error-free HR. The processes that BLM is involved in are indicated by the blue hexagon. Other RecQ helicases involved in the same process are shown by colored hexagons.

The other aspect of BLM's function in DNA replication is telomere maintenance [60–66], and cells deficient in BLM are defective in telomere replication. Telomeres are composed of repetitive nucleotide sequences (in human TTAGGG). One role of BLM is to unwind G4 structures potentially formed by these G-rich repeats. Notably, due to the end-replication problem, telomeres usually erode during cell divisions if there is no intervention. The distal end of a telomere forms a T-loop structure and at the very end of the T-loop, the 3' single-stranded telomere overhang wraps around and invades the double-stranded telomere naturally forming an intertelomeric D-loop (**Fig. 4**). To prevent the DNA ends being recognized by DNA repair machinery, Shelterin complex binds to and stabilizes the T-loop. BLM, as well as WRN, has been found to physically interact with Shelterin proteins [5,62,64]. Resolving the T-loop and D-loop structures by BLM is important to ensure proper telomere replication and telomere length maintenance [60]. Failure to resolve T-loop can cause its cleavage and dramatic telomere shortening. Besides the canonical telomerase-dependent pathway that counteracts the shortening of telomeres, in cancer cells there is a telomerase-independent but HR-dependent mechanism called alternative lengthening of telomere (ALT) in which an adjacent chromosomal telomere (can be sister chromatid) is used as the template to synthesize telomeric DNA. In another model of telomeric extension called rolling circle, cleaved circular telomeric DNA is used as a template to elongate

telomeric DNA. Moreover, the homolog of human BLM in yeast *S. cerevisiae* Sgs1 has been shown to be essential for the yeast equivalent of ALT [60]. Thus, BLM exhibits a division of function at telomeres to suppress HR at telomeres by dismantling T-loop and D-loop in normal cells or facilitating ALT pathways in cancer cells given its roles in HR [60].

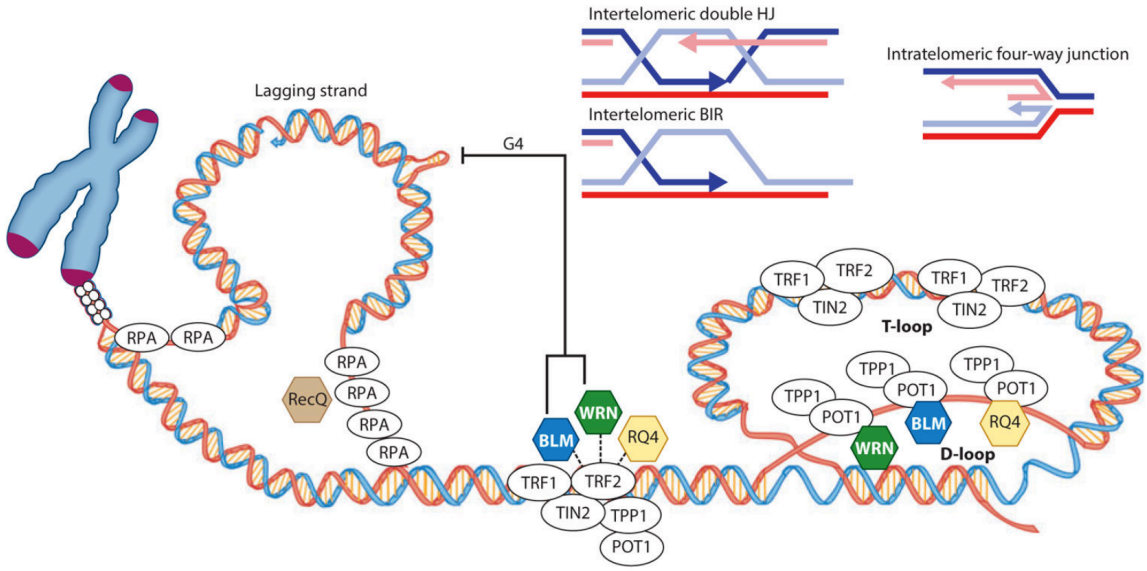



**Figure 4. The role of BLM in telomere maintenance.** BLM and other RecQ helicases may unwind DNA structures such as G4 and D-loops which are potential barriers to telomere replication. BLM physically interacts with components of Shelterin complex composing of Telomere Repeat Binding Factor 1 (TRF1), TRF2, Repressor / Activator Protein 1 (RAP1), TRF1- and TRF2-Interacting Nuclear Protein 2 (TIN2), Telomere Protection Protein 1 (POT1) and Tripeptidyl Peptidase 1 (TPP1), which stabilizes and protects telomere ends. This figure is taken from [5].

## 3.5. The function of BLM in DNA double-strand break repair

Endogenous DSBs occur spontaneously in mitosis (arising from exogenous sources such as ionizing radiation or chemical exposure) or in a programmed manner in meiosis. They are the most deleterious DNA damage among various types of DNA lesions. Human cells can repair DSBs by non-homologous end joining (NHEJ), microhomology-mediated end joining (MMEJ), single-strand annealing (SSA), or homologous recombination (HR) pathways (**Fig. 5**). Choices of different repair pathways depend on the stage of cells and also the end resection of DSB ends. Cells in the S and G2 phases preferentially use error-free HR repair pathways. NHEJ is active throughout the cell cycle though it is imprecise. BLM is promptly recruited to

damaged DNA [49,67] to initiate downstream DNA repair pathways. But unlike WRN, which is involved in all DSB repair pathways, BLM plays a regulatory role only in several steps in HR-dependent DSB repair [4,68].

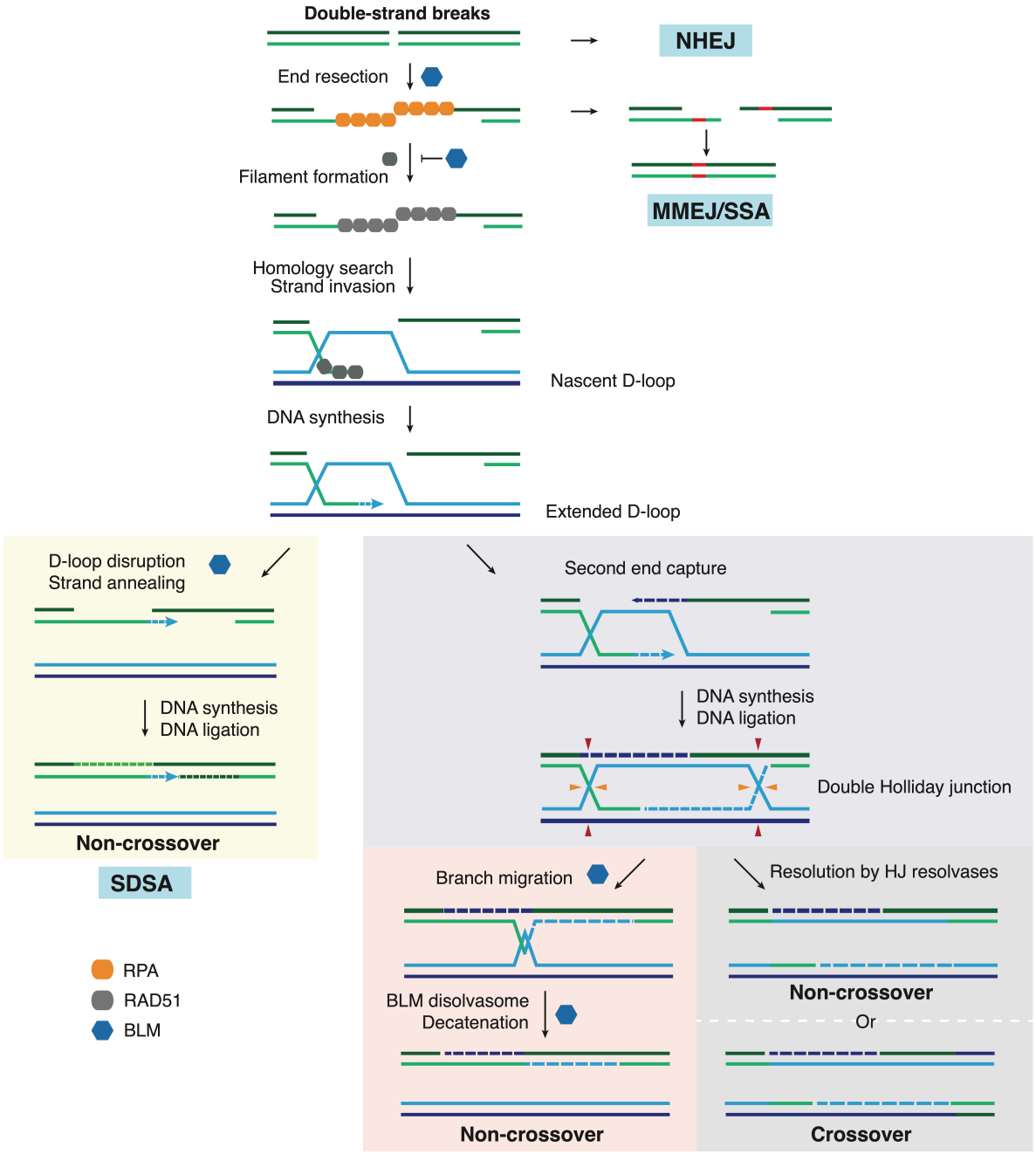



**Figure 5. Models for the repair of DNA double-strand breaks in human cells.** DSBs can be repaired by directly joining the ends together via the non-homologous end joining (NHEJ) pathway. The microhomology-mediated end joining (MMEJ) and single-strand annealing (SSA) pathways rely on homologous sequences flanking the broken junction. Red sequences indicate homologous sequences near the DSB ends. MMEJ depends on microhomology sequences (5-25 bp) while SSA uses long homologies (longer than 30 bp). Homologous

sequences anneal together and the 3' flaps are removed. Subsequently, the gaps were filled by DNA synthesis and DNA nicks are ligated. Either of the aforementioned pathways is error-prone and can lead to deletions. In contrast, HR-dependent pathways repair the DSBs more accurately. After extensive end resection by end-resection machinery containing BLM, followed by the formation of RAD51 filament and D-loop extension, two pathways can be initiated. In synthesis-dependent strand annealing (SDSA) pathway, the extended nascent strand pairs with the other 3' single-strand resected DNA end resulting in disruption of the D-loop. Through further DNA synthesis and ligation, SDSA generates non-crossover products. If the second end is captured, double-HJs will be generated. DHJs can be resolved symmetrically (both junctions cleaved in orientations of red arrows or orange arrows) to generate non-crossover or asymmetrically (one cleaved in orientations of red arrows and the other one in orange arrows) leading to crossover products. Alternatively, dHJs can also be dissolved by BLM dissolvasome generating non-crossover products.

In the HR-dependent DSB repair pathways, the first step is to resect the broken DNA ends to generate a 3' ssDNA overhang. In human, there are two DNA end resection machinery. One consists of BLM, DNA replication nuclease 2 (DNA2), replication protein A (RPA) and MRN (MRE11-RAD50-NBS) complex, meanwhile the other includes Exonuclease 1 (EXO1) instead of DNA2 nuclease [69–71]. Initially, the MRN complex binds to the DSB and recruits RB binding protein 8 (RBBP8, aka. CTBP-interacting protein) to form a primary resection complex. In the first stage of end resection, the MRN complex primarily resects the ends. Subsequently, the MRN complex recruits either BLM-DNA2 or BLM-EXO1 to the ends to initiate extensive end processing. RPA then binds to and protects the single-stranded DNA (ssDNA) ends. In the next step, RAD51 binds to the resected ssDNA and forms right-handed helical filaments on it. Subsequently, the RAD51-ssDNA filament searches for homologous sequences and forms a nascent D-loop by strand invasion. Notably, BLM can inhibit D-loop formation by inhibiting the nucleation of RAD51 [32].

After priming DNA synthesis and D-loop extension, two DNA repair pathways can subsequently be initiated. If the D-loop is disrupted through the annealing of the newly synthesized DNA strand to the other break end in its complementary strand, double-strand breaks (DSBs) are repaired by synthesis-dependent strand annealing (SDSA). The homologous recombination (HR) process is then completed by gap repair via DNA synthesis followed by ligation (**yellow area in Fig. 5**) [72–75]. Notably, the SDSA repair pathway always generates non-crossover products, thus avoiding crossovers and loss

of heterozygosity (LOH) [72]. Therefore, this mechanism is normally used in somatic cells [72,76]. Alternative to SDSA, once the D-loop is extended, it can capture the second end and form a dHJ intermediate. dHJs can be dissolved by BLM dissolvasome (aka. BLM/Topo IIIα/RMI1/RMI2 complex or BTR complex; RMI, RecQ-mediated genome instability; Topo IIIα, topoisomerase IIIα), thereby generating non-crossover products. BLM promotes the two junctions migrating towards each other and in the end, BLM dissolvasome decatenates the strands of the hemicatenane (the intermediate product when the two HJs migrate towards and eventually meet each other) assuring non-crossover products (**Fig. 5, pink area**). DHJs can also be processed by Holliday junction resolvases, GEN1 Holliday junction 5' flap endonuclease and MUS81/EME1 complex [77]. If the two junctions are resolved symmetrically (they are cleaved in the same orientation), non-crossover products will be generated. However, the asymmetrical resolution of the junction (cleavage in different orientations) leads to crossover products (**Fig. 5, grey area**) [8,77,78].

In the NHEJ pathway, unresected or minimally resected break ends are directly ligated with no need for a homologous template. As a result, this imprecise repair may lead to the loss of a few nucleotides. When the NHEJ pathway is inactivated, DSBs can be repaired microhomology-mediated end joining (MMEJ, also known as alternative NHEJ), which relies on microhomology sequences near the break ends. Microhomology sequences anneal after end resection, and any extra 3’ tail is subsequently trimmed. Following DNA synthesis to fill in the gaps and DNA ligation, DSBs are repaired though, it often results in DNA deletion.

The steps of DSB repair by single-strand annealing (SSA) are very similar to MMEJ. Unlike MMEJ, SSA requires the annealing of longer near the DSB ends, typically over 30 bp [5,7,72]. Therefore, SSA repair may result in large DNA fragment deletion and is also mutagenic. Choices of different repair pathways depend on the stage of cells and also the end resection of DSB ends. NHEJ is active throughout the cell cycle though it is imprecise. Cells in S and G2 phases preferentially use error-free HR repair pathways. It has been shown that depletion of BLM in cells stimulated HR repair.

## 3.6. BLM and recombination

DNA replication, DNA repair, and DNA recombination are interlinked processes. The

definition of recombination is inconsistent among molecular biologists, evolutionary biologists, geneticists and biochemists. Here, to avoid any misunderstandings, a definition in the context of the molecular processes and molecular outcomes is used, that is, genetic recombination refers to the rearrangement of DNA sequences by some combination of processes of breakage, rejoining, and copying of chromosomes or chromosome segments [79,80]. BLM has a pro-recombinant role given that BLM acts in the early key step of end resection in HR, [70,81]. On the other hand, BLM can disrupt D-loops and BLM dissolvasome dissolves dHJ generating non-crossover products thus suppressing recombination [8,32,55]. Notably, dissolving dHJs is a function unique to BLM and distinguishes it from other helicases.

Depending on when the recombination occurs, DNA recombination in the living organism can be divided into mitotic and meiotic recombination [7,82,83]. Both mitotic (e.g., SCEs, LOH) and meiotic recombination are results of DSBs followed by HR in which BLM has indispensable roles [5,82,83]. In mitosis, BLM deficiency leads to excessive SCEs and elevated frequency of LOH [14,84]. In meiosis, it has been observed in zebrafish, drosophila and mice that BLM deficiency leads to a strikingly increased number of crossovers, defective homologous chromosome pairing and abnormal chromosome segregation during meiosis or cell apoptosis [85–87]. This feature of elevated recombination during mitosis has been leveraged to study and compare gene functions [88,89].

Altogether, BLM is crucial to suppress inappropriate recombination during mitosis and meiosis and it displays both pro- and anti-recombination roles.

# 4. Bloom Syndrome

Bloom Syndrome (BS), also termed congenital telangiectatic erythema and stunted growth, is a rare autosomal recessive disorder caused by loss-of-function mutations in *BLM* gene [1].

## 4.1. Bloom-Syndrome causing mutations

Overall, 105 mutations in *BLM* gene have been reported worldwide, some of which are identified in individuals with cancers instead of BS (https://www.LOVD.nl/BLM).

Among the 250 cases registered in the Bloom's Syndrome Registry, 64 syndrome-causing mutations have been characterized with 54 of them resulting in premature protein-translation termination and the rest 10 being missense mutations (**Fig. 6**) [44]. Premature protein-translation termination mutations spread all over the genes whereas missense mutations were only found in the helicase domain. German et al. characterized the molecular features of the BLM mutants *in vitro* [90]. One feature of the mutations is that they are founder mutations, in that many of these mutations are found at unusually high frequencies in (isolated) populations. A concrete example is that BS is more common in individuals of Ashkenazi Jewish descent than in any other population. The predominant BS–causing mutation*, BLM* c.2207-2212delATCTGAinsTAGATTC Ashkenazi Jewish BS mutation (*blm*$^{Ash}$), is homozygous in nearly all Ashkenazi Jewish individuals with BS. Ellis et al. found that the *blm*$^{Ash}$ mutation is present in 1 of 107 of this particular Ashkenazi Jewish population or a carrier frequency of 0.0093 whereas the carrier frequency in the normal population is extremely low (no available precise number) [91]. More than half of the affected cases are homozygous of their BS-causing mutation; some are genetic compounds, in which two *BLM* alleles carry different mutations; and in some cases, only the pathogenic mutation in one *BLM* allele was characterized. Mutational analysis of samples directly from BS individuals and cell lines with stably integrated *BLM* mutations showed that all the protein truncation mutations lead to a striking reduction in *BLM* mRNA, which was probably due to nonsense-mediated mRNA decay (NMD), and eventually result in the dysfunction of BLM protein in cells [90]. The expression of BLM protein in some cell lines derived from individuals with BS (i.e. GM08505, GM03403D) were undetectable [17]. All ten putative BS-causing missense mutations occur at highly conserved residues within the helicase domain. *In vitro* studies of the molecular effects at the protein level of missense mutations, proved that mutations in murine Blm protein homologous to Q672R, I841T and C1055S in human BLM protein impair or even abolish the catalytic activities of Blm [43,44,92]. Further studies provided structural insights that Q672R, I841T, G891E and C901Y mutations affect the function of the protein due to the changes in the local electrostatic environment and local structures [93]. BS-causing evidence for missense mutations also came from the fact that transfecting *BLM* cDNA with these missense mutations into BS cells could not correct the hyper-

recombination phenotype in BS cells [44,94]. It is noteworthy that BS carriers are at an increased risk of colorectal cancer even though BS carriers do not show BS phenotypes [95]. Moreover, in addition to these BS-causing mutations, Alzahrani et al. employed a computational approach to systematically evaluate the pathogenic risks of SNPs in the *BLM* gene [96].

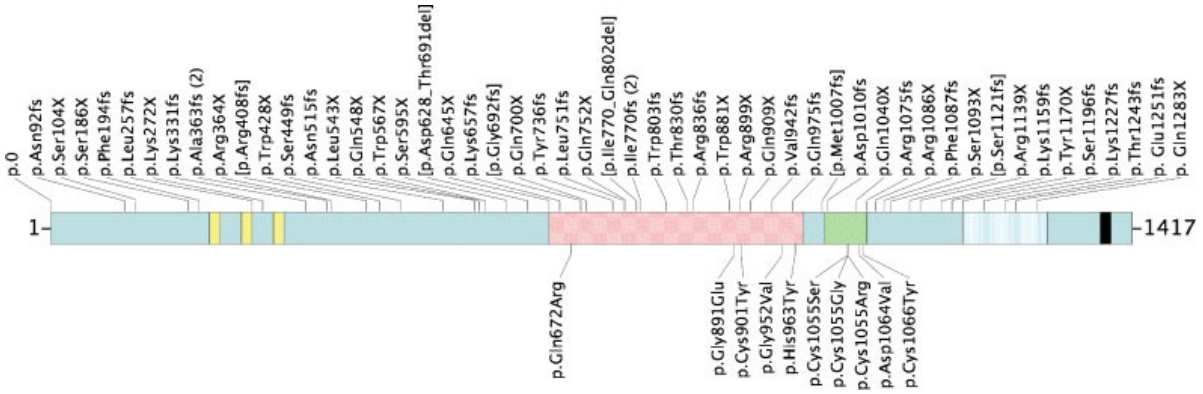



**Figure 6. 64 identified mutations of *BLM* gene in BS patients in Bloom Syndrome Registry.** The colored boxes depict the full-length BLM protein. The mutations shown above the diagram putatively result in premature protein-translation termination, and mutations below the diagram lead to amino acid substitutions, all of which occur within the helicase domain, indicated by the pink and green regions. This figure is taken from [44].

## 4.2. Clinical and molecular phenotypes of Bloom Syndrome

Clinical phenotypes of BS are multifaceted and most patients die before the age of 30 [9,90]. Notably, premature aging is not a trait of BS [97]. The most prominent and the only consistent clinical feature is dwarfism due to pre- and postnatal growth retardation, e.g., disproportionately small head. Most patients have a strikingly low birth weight [98]. According to the data from the Bloom Syndrome registry (https://pediatrics.weill.cornell.edu/research/bloom-syndrome-registry), the mean height of adult patients of BS is 149 cm (range from 128 cm to 164 cm) for men and 138 cm (range from 115 cm to 160 cm) for women (Bloom Syndrome registry). They are prone to develop chronic obstructive lung diseases, diabetes, and cancer prematurely. Cancer is the most frequently observed complication in individuals with BS and is the leading cause of death. 122 out of 256 people in the registry developed cancer at a mean age of 26. The cancer types are not limited to certain types but they

are as diverse as in the general population. The most frequent types of cancer in BS are lymphoma and leukemia. The mean age at which acute myelocytic leukemia (AML) or acute lymphocytic leukemia (ALL) are diagnosed is 18 (range 2 - 47) and 20 (range 5 - 40) respectively [12,98–100].

BS patients also often suffer from reduced fertility [9,101]. Male patients are infertile and found to be azoospermic while female patients are subfertile [90,99,102–104]. Puberty in females is delayed and most of the fertile female individuals of BS experienced early menopause [99]. Cases of women with successful delivery have been reported [99]. Other main clinical features include (1) feeding difficulties in infancy, the reason for which is still unclear; (2) abnormal facial morphology with a narrow face and protruding ears; (3) sun-sensitive erythema affecting the butterfly area of the face and sometimes hands and forearms; (4) spots of hyper- and hypopigmentation of the skin, described as “café au lait” spots by German; (5) immune deficiency; (6) frequent infection especially of the middle ear and lung; (7) high-pitched voice; (8) normal intelligence or mild mental retardation [90,98]. At the cellular level, chromosomal abnormalities are hallmarks of BS and BLM deficiency gives rise to a characteristic phenotype of a high frequency of sister chromatid exchange and genomic instability, which contributes to the predisposition of cancer in individuals with BS [9,11,13,90].

So far there are no clinical diagnostic criteria for BS. The diagnosis is usually made after some features of abnormal growth and medical problems are recognized and is further confirmed by increased frequencies of sister chromatid exchanges in lymphocytes and fibroblasts from patients with BS (**Fig. 7**) [13]. The cellular hallmark of BS is hyper-recombination. German et al. studied the sister chromatin exchange (SCE) in *in vitro* culture of lymphocytes from BS patients, BS carriers and non-BS donors. They observed that the SCE frequency is increased by 12 folds in BS (**Fig. 7a, 7b**). The mean of SCEs per cell at metaphase is 89 (range 45 to 162) in lymphocytes from individuals of BS, 6.9 (range 1 to 14) in cells from carriers, and 8.2 (range 2 to 19) in cells from healthy donors [12,13]. In line with cytogenetic studies, Van Wietmarschen et al. used single-cell sequencing techniques to map the SCEs in BLM-deficient cells at single cell level and confirmed the 7 ~ 10 10-fold increase in the frequency of SCEs [25].

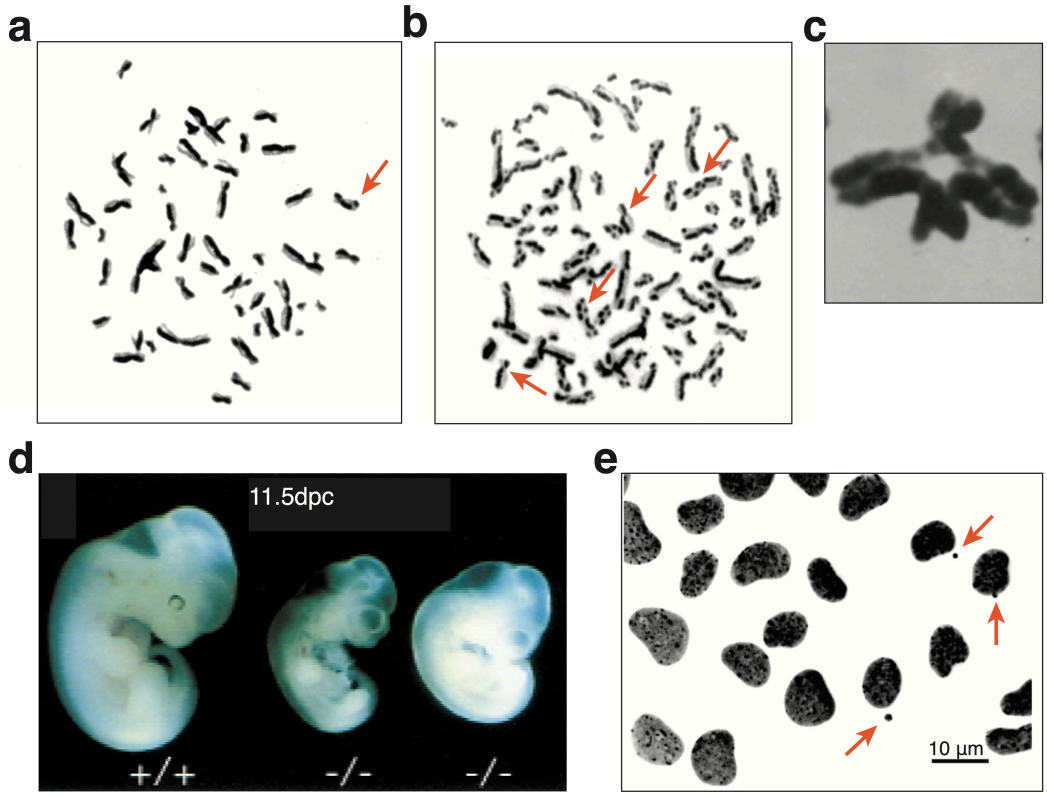



**Figure 7. Analysis of chromosomal instability in cells from individuals of BS.** (**a**) The sister chromatid exchanges (SCE) in a normal lymphocyte at the second metaphase. Arrows indicate the breaking points of exchanges between SCE [99]. (**b**) The sister chromatid exchanges (SCE) in a BS lymphocyte at the second metaphase. An elevated SCE frequency can be observed (enlarged 1800x) [99]. (**c**) Chromatid interchange configuration composed of two probably homologous chromosomes from a Bloom Syndrome lymphocyte (enlarged 4150X) [13]. (**d**) Normal 11.5-dpc (days post coitum) *Blm*$^{+/+}$ and mutant *Blm*$^{-/-}$ embryos [10]. (**e**) Murine *Blm*-deficient cell lines with micronuclei indicated by arrows [16]. Figures are taken from the corresponding cited articles.

Exchanges between both sister chromatids and non-sister homologous chromatids generating quadriradial configurations were also observed in BS lymphocytes (**Fig. 7c**) [13,105,106]. Additional chromosomal abnormalities associated with BS include an increase in DSBs, translocations, and the formation of chromosomal radials [13]. Consistent with the observation, LaRocque et al. showed that there was a five to ten-fold increase of loss of heterozygosity (LOH) in BLM-deficient cells by tracking certain gene markers [14]. LOH can arise from HR-dependent pathways as well. While SCEs arise from mitotic recombination between identical sister chromatids and are innocuous, LOH is due to mitotic recombination between homologous chromosomes which do not necessarily contain identical genetic information. LOH can also arise from

NHEJ. NHEJ is an error-prone pathway in DNA DSB repair and has been observed at higher levels in BS cells [107]. In human, LOH in tumor suppressor genes is a critical event during the development and progression of cancers [108,109] and the increased frequency of LOH presumably contributes to cancer predisposition in BS [13]. As described earlier, NHEJ pathway is an error-prone repair pathway and is active through the entire cell cycle. NHEJ cannot be initiated from free 3' ssDNA after end resections. BLM is involved in 5' end resection in the early steps of HR and this activity shuffles the choices of DSB repair pathways away from NHEJ towards other more accurate DNA repair pathways. Nevertheless, using single-cell template strand sequencing (Strand-seq) instead of gene markers to track LOH, a recent study showed results that were contradictory to what has been previously demonstrated – increased frequency of LOH in BS cells [14,25]. Instead, they found that the frequency of LOH is not significantly elevated in BLM-deficient cells although slightly more LOH regions were detected in BLM-deficient cells [104]. Moreover, many mouse models have been generated to study the function of BLM. Originally, *Blm* was thought to be an essential and indispensable gene in mice as homozygous *Blm* knockout mice are not viable and embryonic lethal by day post coitum (dpc) 13.5 [10]. They exhibited many clinical and molecular manifestations of BS and genome instability. Apart from elevated frequency of SCEs in embryonic fibroblasts from these embryos and micronuclei in isolated red blood progenitor cells, mouse embryos showed a small size and severe developmental delay (**Fig. 7d**), mimicking proportional dwarfism seen in BS patients [9,10]. However, later Luo et al. generated a viable *Blm* mutant mouse carrying two different mutant *Blm* alleles [11]. They observed the same increased mitotic recombination phenotypes including SCEs and somatic LOH in Bloom mice as seen in BS patients [11]. They demonstrated that LOH from increased frequency of somatic recombination constitutes an underlying mechanism of tumor susceptibility in these mice [11].

Another piece of evidence for chromosome instability *in vivo* in BS patients was that fibroblast cells and exfoliated cells obtained from the oral cavity and the urinary tract from BS patients showed strikingly higher frequencies of cells with micronuclei (**Fig. 7e**) [111,112]. Micronuclei are the chromosome fragments or lagging whole chromosomes not included in the daughter nuclei produced by mitosis because they fail to correctly

attach to the spindle during the segregation of chromosomes in anaphase and in turn generate an extra micronucleus in daughter cells alongside the nucleus [113,114]. Micronuclei may arise from the disrepair of double-strand breaks if DNA damage exceeds the repair capacity of cells and incorrect repair of DNA breaks in the case where the HR repair pathway is dysfunctional [113]. Of note, fibroblasts from BS carriers exhibit increased micronucleus formation as well, recalling the increased risk for cancer in heterozygous carriers [95,112]. Additionally, BLM-deficient cells have higher mutation rates [111]. Other molecular phenotypes seen in BS cells include a slower rate of replication fork movement and a longer S phase [17,59]. To summarize, BLM-deficient cells exhibit a wide range of features indicative of genomic instability and are prone to acquire mutations.

Furthermore, BS cells display altered gene expression profiles. The mRNA and miRNA gene expression profiles were altered in fibroblasts from BS patients in comparison to matched healthy donors. There was a significant enrichment of G4 motifs at transcription start sites, and especially with the first intron of BLM-regulated genes. Top enriched molecular and cellular functions of BLM-regulated genes included “cell growth/proliferation,” “cell death/survival,” “protein synthesis,” “gene expression,” and “cellular development,” with “molecular mechanisms of cancer” being the top-ranked canonical pathway. Regarding more detailed molecular functions, they were enriched in Notch signaling pathways, TGF-β and NF-κB signaling, protein kinase activity, and nucleotide binding and nucleic acid metabolic processes. Many of BLM-regulated genes associate with diverse immunologic cell types and processes, which is consistent with the data from *BLM*-depleted cell lines. Differentially expressed miRNAs are often cancer-associated [18,26]. Altogether, these provide further evidence to correlate the absence of BLM and cancer-predisposition observed in BS.

## 5. G-quadruplexes

G-quadruplexes (G4) are four-stranded non-canonical structures formed by single-stranded guanine-rich sequences [34,109]. Four guanine bases use G-G base pairing through Hoogsteen hydrogen bonding and form a planar structure called G-quartet; then two or three G-quartets stack on each other forming a G4 (**Fig. 8a**). G-

quadruplexes can be formed by DNA, RNA and DNA-RNA hybrid molecules [110,111]. Depending on the nucleic acids strand orientation, G4s have diverse topologies (**Fig. 8b**). Moreover, G4s can be formed within the same molecule (intermolecular) or among multiple molecules (intramolecular) (**Fig. 8b, 8c)**. In this thesis, mainly DNA G-quadruplexes are discussed and for convenience, G-quadruplexes and G4s refer to DNA G-quadruplex structures if not specified otherwise. G4s were first proposed in 1962 and initially observed in telomeric DNA oligonucleotides *in vitro* [109]. Over the last few decades, significant strides have been made in our understanding of G4s by characterizing their structures, biophysical properties, their distribution in the genome, small-molecule G4 stabilizers and destabilizers, G4-unwinding helicases, G4-specific antibodies, G4-interacting proteins, and importantly, their biological functions [20,34,45,110,112–119]. Recent advancements in techniques have enabled us to visualize and even track their dynamics *in vivo* with live imaging. G4 (motifs) have been implicated in developmental processes and diseases, such as cancer and Bloom Syndrome, and are considered promising therapeutic targets in cancer [110,111,120].

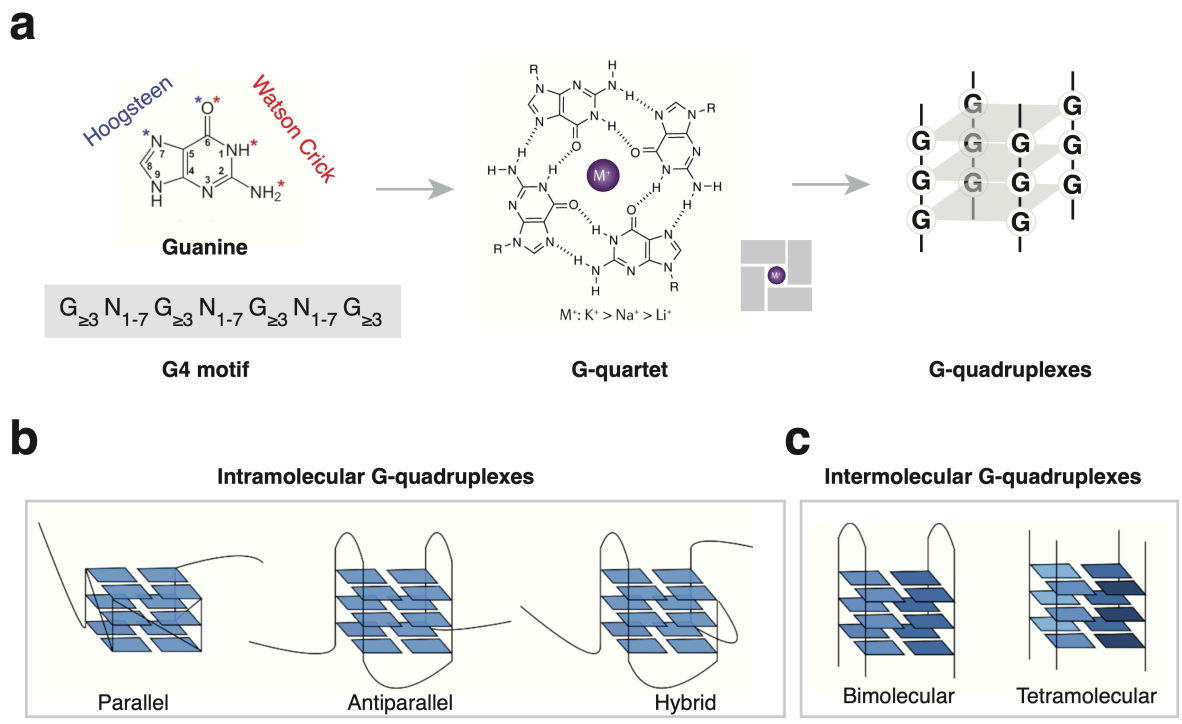



**Figure 8. Structure of G-quadruplex.** (**a**) From Guanine to G-quadruplexes. Four Guanines form a planar G-quartet and two or three G-quartets form a G-quadruplex. (**b**) Topologies of various intramolecular G-quadruplexes according to the direction of the DNA backbones - parallel, antiparallel and hybrid. (**c**) Example of intermolecular G-quadruplexes formed among two or four DNA molecules. Figures adapted from [19,127].

## 5.1. Mapping G4 (motifs) in the human genome

### 5.1.1. *In silico* approaches

The propensity to fold into G-quadruplexes can be predicted from the primary DNA sequences. The canonical putative G-quadruplex sequences, or G4 motifs, are $G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}$. A convenient and the most straightforward approach to identify them is to use bioinformatic tools to scan the genome for specific patterns [19]. The sequence composition of G4 motifs, especially the non-canonical G4 motifs, largely depends on the knowledge from *in vitro* experiments. To date, there are various tools available with different underlying algorithms though, most of the existing tools are limited in detecting intramolecular G4 motifs. There are also a few tools available to detect intermolecular G4 motifs, e.g., *AllQuads*. The number of detected G4 motifs in the human genome depends on both the motif pattern and the bioinformatic tools [19,20]. Using the canonical G4 motif pattern, about 400,000 to more than 1000,000 G4 motifs can be identified in the human genome depending on the bioinformatic tools [19,110]. Putative G4-forming sequences matching certain patterns in the DNA sequence predicted by bioinformatic tools are herein referred to as G4 motifs.

### 5.1.2. *In vitro* approach by G4-seq

G4-seq is a method developed to map DNA sequences that can form G4 structures *in vitro* across the genome (**Fig. 9a**), which takes advantage of the fact that the formation of G4 blocks DNA polymerase [20]. The genomic DNA is first sequenced under condition disfavoring G4 on an Illumina flow cell (Read 1); subsequently, the synthesized strand during Read 1 is peeled away and the same DNA is re-sequenced under conditions favoring G4 (Read 2). During the sequencing process for read 2, when the polymerase reaches G4s, the polymerase will be stalled and start adding gibberish bases, which cause a sudden drop or discontinuities in base quality. By comparing sequences and their qualities in read 1 and read 2, G4-forming sites can be inferred. With this approach, 716,000 G4 structures were identified in the human genome and more than half of them were not canonical G4 motifs predicted by computational methods [20]. Putative G4-forming sequences identified by G4-seq are herein referred to as G4-seq hits.

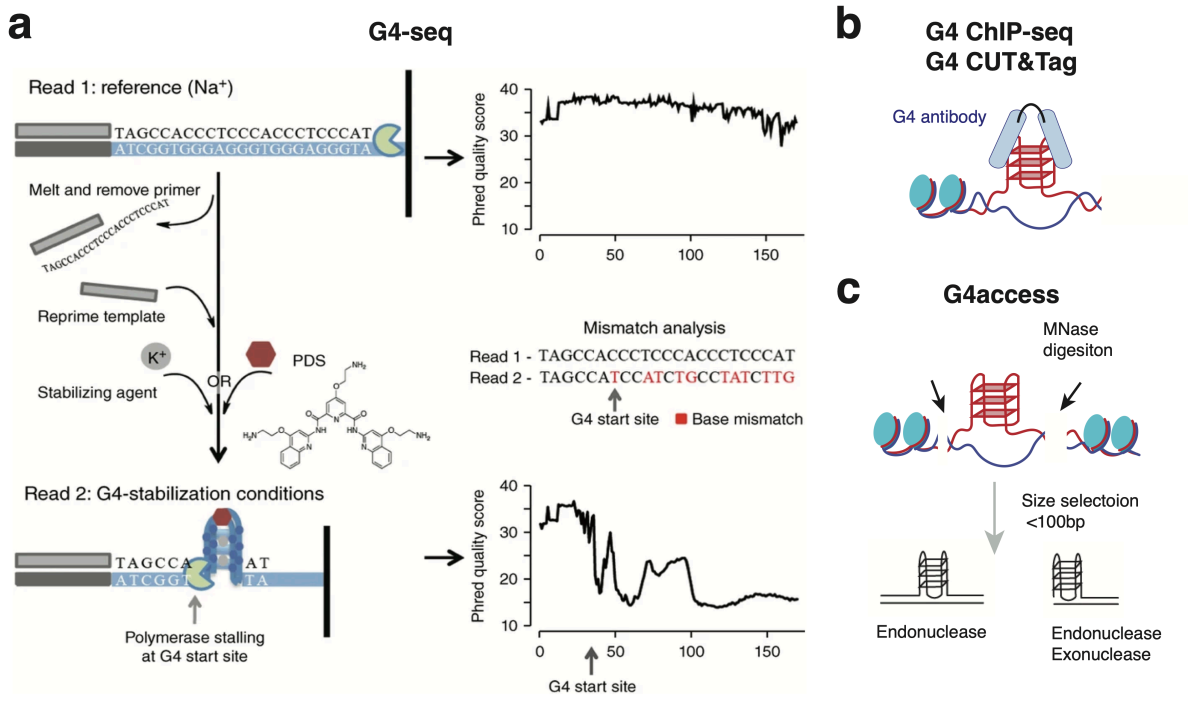



**Figure 9. Methods to map G4 genome-wide.** (**a**) The principle of G4-seq, which takes advantage of G4-induced polymerase stalling leading to a sharp drop in sequencing quality. Read 1 is sequenced in G4-diafavoring condition and read 2 is sequenced in G4-favoring and G4-stabilizing condition. PDS, pyridostatin, a small molecule that stabilizes G4. (**b**) Antibody-based methods, G4 ChIP-seq and G4 CUT&Tag. They utilize specific antibodies (e.g., BG4) that recognize G4 structures. (**c**) The concept of G4access–the structure of G-quadruplexes protects the focal DNA from being digested by micrococcal nuclease (MNase), exonuclease and endonuclease. Figures adapted from [20,119].

### 5.1.3. Methods to detect endogenous G-quadruplexes

Compelling evidence has successfully demonstrated the existence of G4 structures in vivo. One common concept in such approaches is to capture G4s via G4-specific antibodies and map their position via sequencing, which was made possible by the recent advancement in G4-specific antibody and genomic techniques. Clone 1H6 [124], which is a monoclonal antibody generated using vertebrate and ciliate telomeric G4s as immunogens, was used frequently in the past but has now been largely replaced by BG4, an antibody with higher specificity [123]. BG4 was screened and isolated by phage display and it binds to various G-quadruplex conformations *in vitro* [123]. BG4 has been used to develop G4 ChIP-seq and G4 CUT&Tag which map genome-wide endogenous G4s (**Fig. 9b)** [21,24]. G4 ChIP-seq adapts the traditional ChIP-seq method. It requires a relatively large amount of material (several tens of millions of cells).

Chromatin is fixed and sheared into small fragments and subsequently chromatin fragments with G4 structure are enriched [21]. G4 CUT&Tag is a recently developed alternative to ChIP-seq, which is simple and improves signal-to-noise ratio in the data [24]. It requires much lower input material (~100,000 cells) [24,128]. A critical difference from G4 ChIP-seq is that it does not require chromatin fragmentation by sonication. Instead, the antibody is incubated with intact permeabilized cells and binds to the target, which avoids potential biases and loss of G4s in heterochromatin regions; then cells were incubated with secondary antibody, and protein A-fused Tn5 transposome; subsequently via protein A or secondary antibody binding to BG4 antibody, Tn5 is tethered to G4s; Tn5 is activated and inserts the sequencing adapters into the chromatin flanking G4 sites upon the addition of $Mg^{2+}$; lastly, in PCR step, DNA fragments containing G4 are predominantly amplified [24,128]. Notably, G4 CUT&Tag can be performed on native and unfixed samples and adapted for single-cell assays. Only about 10,000 to 20,000 G4-forming sites were identified in different cell lines, though the number is higher in immortalized cell lines compared to primary cell lines [21,23,24]. Meanwhile, Esnault et al. developed an antibody- and crosslinking-independent method, G4access (**Fig. 9c**) [119]. It utilizes Micrococcal Nuclease (MNase) to digest chromatin and in the sub-nucleosomal DNA fraction, DNA with G4 formation is protected from being digested [119]. This technique is convenient and not biased by the affinity and/or specificity of the G4 antibody recognizing G4 structures, albeit it may capture DNA with other alternative structures or bound by proteins.

## 5.2. Biological functions of G-quadruplexes

G4 motifs are abundant in the telomeres and are highly enriched in promoters and enhancers and most of the G4s form in the open chromatin region and reside in promoters, suggesting the broad regulatory potential. In the last few decades, emerging studies have been carried out and unraveled their regulatory functions in various biological processes under physiological conditions (**Fig. 10**).

### 5.2.1. DNA replication

The formation of G-quadruplexes during replication may present a barrier to the DNA replication machinery, impeding the progression of DNA polymerases and initiating DNA repair pathways to reset the replication fork (**Fig. 10a**) [116,129]. However, they have

also been shown to stimulate the initiation of DNA replication. In *Drosophila*, mouse, and human cells, more than 60% of DNA replication origins contain the origin G-rich Repeated Element (OGRE), which can potentially form G4s [130–133]. Prorok et al demonstrated that OGRE/G4s ware functionally required for the initiation of DNA replication and they found that the deletion of the OGRE/G4 sequence strikingly decreased corresponding origin activity whereas inserting such an element created a new replication origin [134].

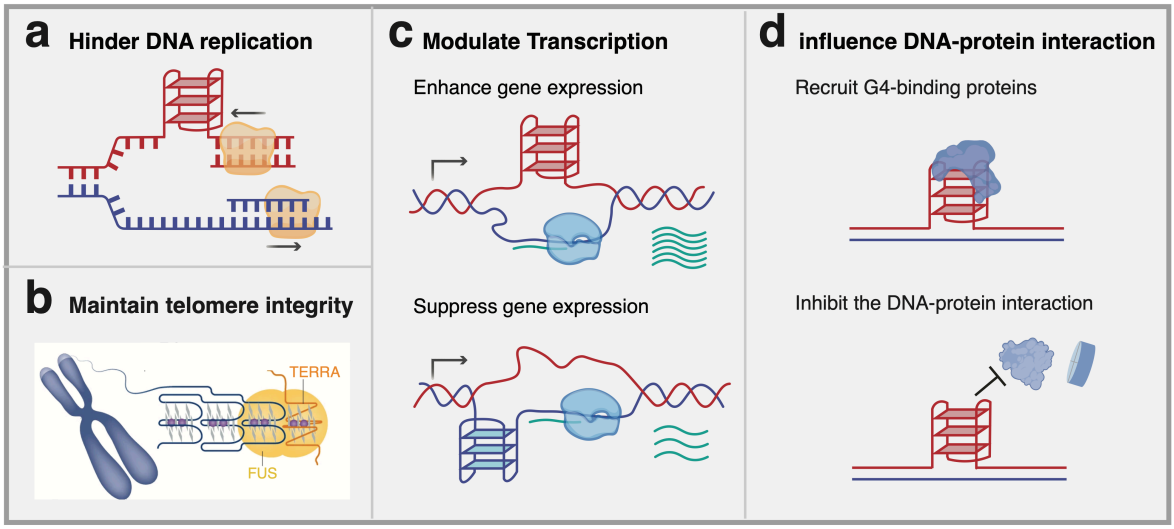



**Figure 10. Regulatory function of DNA G-quadruplexes.** (**a**) The formation of G-quadruplexes hinders the progression of DNA replication machinery. (**b**) The formation of G-quadruplexes in telomeres stabilizes the telomeres. (**c**) G-quadruplex formation in promoter enhancing and repressing transcription. (**d**) G-quadruplex formation influences DNA-protein interaction, e.g., recruiting G4-binding proteins or displacing histones. The panel **b** is adapted from [127].

### 5.2.2. Telomere maintenance

A telomere is a specific region in eukaryotes at the end of the chromosome consisting of species-specific repetitive DNA sequences (TTAGGG in vertebrates) and proteins [135]. To prevent the chromosomal ends from being recognized as DNA double-strand breaks by DNA repair systems, telomeric ends usually form a T-loop structure which is bound by Shelterin complex. G-rich sequences at the telomere provide a favorable niche for G4s to form. Using immune-staining with BG4, G4s have been directly observed to form at telomeres [123]. It has been shown in ciliates that instead of T-loops, the very end of the telomere forms G4s to stabilize the telomere (**Fig. 10b**) [136,137].

However, the formation of high levels of G4 might impose great stress on DNA replication machinery and could lead to telomere shortening (**Fig. 10b**) [116]. This is in accordance with the telomere shortening and fragility observed in cells lacking G4 unwinding helicases BLM, WRN, and RTEL1 [138–140]. Some of the proteins in Shelterin complex can recognize G4s [141,142]. In short, the formation of G4s is a double-edged sword at the telomere.

### 5.2.3. Transcription

The role of G4s in transcription is one of the most extensively studied aspects. Endogenous G4s detected by G4 ChIP-seq are mostly present in promoters [22–24]. Particularly, many oncogenes, like *c-KIT*, *c-MYC*, *KRAS* harbor G4 motifs in their promoter. It has been proposed that the formation of G4 can hinder the loading of RNA polymerase onto the DNA or impede the progression of RNA polymerase, thereby inhibiting gene transcription, and vice versa facilitating it (**Fig. 10c**) [34,110,111].

The first evidence of G4s' regulatory functions came from the observation in *MYC* gene that stabilizing the G4 in the nuclease hypersensitivity element $III_1$ upstream of the P1 promoter represses its transcription [137]. A recent study on the G4 motif in *MYC* gene suggests that G4 formation is a positive regulator of *MYC* transcription and comprehensively investigated the molecular mechanisms by which G4s directly and indirectly modulate transcription. Esain-Garcia et al. abrogated the G4-forming potential via genome editing and profiled a few key molecular signatures for transcription [138]. They showed that upon the loss of G4-folding, there is lower transcription, decreased binding of transcription factor SP1 and CNBP1, nearly abolished signal of histone methyltransferase, MLL1 and MLL4 and lower levels of H3K3me1 and H3K4me3. Moreover, at the mutant G4 locus, they observed a nascent nucleosome and significantly decreased RNA polymerase II occupancy in the *MYC* promoter [138]. In this case, G4 formation in the *MYC* promoter orchestrates various transcriptional events, including TF binding, histone modification, nucleosome occupancy, and RNAPII positioning. Li et al. demonstrated that enhanced G4 persistence via G4 stabilizing molecule impairs transcription initiation [139].

In addition, G4s can influence transcription via interacting with R-loops, which is a three-stranded structure formed by RNA transcripts invading into, and displacing one

of the strands in DNA. Formation of G4s is associated with the presence of R-loop [139,140]. Lee et al. demonstrated that G4-formation on the non-transcribing strand stabilizes R-loop and promotes transcription by successive R-loop formation [141]. Interestingly, they also found that orientation of G4s in the promoter matters: G4s formation on the non-transcribing strand strongly enhances the gene expression whereas G4 formation on the transcribing strand could completely block transcription [141]. To summarize, the precise downstream impacts on transcription appear to be highly context-dependent.

### 5.2.4. DNA-protein interactions

G4s could influence the interaction between DNA and proteins and subsequently modulate epigenetics and/or transcription. In human, the transcription factor (TF) binding sites are enriched in the G4 motifs and these TFs were shown to recognize and bind to G4 structures [68,82]. Therefore, G4 could recruit TFs to certain genomic sites and modulate transcription (**Fig. 10d**).

As mentioned in the previous section, G4 could interfere with RNA polymerase II loading on the DNA, influencing transcription initiation and elongation. The formation of G4s may displace the histones or hinder histones binding back to local DNA, thereby altering chromatin accessibility (**Fig. 10d**) [22,138,139]. G4s could also spatially block the access of DNA methyltransferase and histone-modifying enzymes to the DNA or chromatin, thereby modulating the epigenetic states [138,143].

Broadly speaking, G-quadruplexes modulate biological processes in two major ways. The first way relies on the spatial structure that G4s act as an obstacle for polymerases, or displace other proteins binding to the local DNA. The second way involves G4 recruiting other factors at specific genomic loci. It has been shown that G4 motifs contain binding sites for a lot of transcription factors in human and some transcription factors have higher binding affinity for a specific conformation [115,142].

# 6. Links between putative G4-forming sequences and Bloom Syndrome

The prevalence of G4-forming sequences was implicated in SCEs, the characteristic feature of BS, and the differentially expressed genes in BLM-deficient cells. Van Wietmarschen et al. performed Strand-seq and characterized SCEs at ~ 10kb resolution in eight cell lines, four derived from healthy donors and four derived from BS individuals. They observed that the SCEs in both WT and BS tend to occur at the common fragile site hotspots despite no stronger enrichment of SCEs from BS cells. However, they observed a remarkable enrichment of G4 motifs in SCE sites in BS and notably not in WT cells, especially when G4 motifs exist on the transcribed strand of the gene [110]. Given BLM's role in unwinding G4 structures *in vivo* and *in vitro* and the numerous prevalence of G4 motifs in the genome, they proposed a mechanism, in which in the absence of BLM, G4 structures are more likely to persist and interfere with DNA replication, which subsequently initiate DNA repair pathways and eventually leads to elevated SCEs together with the lack of BLM to dissolve dHJ [25,73].

Differentially expressed genes were also correlated with the presence of G4 motifs. It was found that G4 motifs were enriched at differentially expressed genes in fibroblasts from individuals with BS and in BLM-depleted control fibroblasts, particularly at the junction site of the first exon and the first intron on the non-transcribing strand [18,26]. Another study, however, argued that this enrichment is due to the bias of GC content in these regions and there was no significant enrichment of G4-forming sequences detected using G4-seq and expanding the sample size [27].

These studies suggest an association of G4 motifs in BS, however, neither *in silico* predicted G4 motifs nor *in vitro* validated G4-seq hits in purified genomic DNA can reliably can reliably represent the G4-forming status *in vivo*. It remains unclear if the endogenous G4 profile is disrupted under the BLM-deficient condition and what regulatory roles, if any, they play in the molecular phenotypes in BS.

# Chapter 1

## G-quadruplexes act cooperatively with open chromatin in mediating gene expression changes in Bloom Syndrome

Dingwen Su [1*], Veronika Altmannova [1], Volker Soltys [1], Moritz Peters [1], Christopher M Cunniff [2], John R Weir [1], Yingguang Frank Chan [1,3*]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany

2 Department of Pediatrics, Weill Cornell Medical College, New York, NY, USA.

3 University of Groningen, Groningen Institute for Evolutionary Life Sciences (GELIFES), 9747 AG Groningen, The Netherlands

* Corresponding authors

### Status in publication process

Advanced manuscript; ready for submission to journals

### Author contributions

D.S. conceived the project, designed the experiment, generated the data, conducted the data analysis, and interpreted the results. Y.F.C. supervised the data collection and helped interpret the results. V.A. provided experimental support and support for the graphics. M.P., V.S., J.W., and Y.F.C. provided experimental or computational support. C.M.C. kindly provided the samples from the Bloom Syndrome family. D.S. wrote the paper with input from all authors. All authors reviewed and approved the final version of the manuscript.

## Abstract

Bloom Syndrome (BS) is a recessive genetic disorder characterized by hyper-recombination and genome instability. It is caused by mutations in the conserved RecQ helicase gene, *BLM*, which is essential in maintaining genome integrity and unwinds various aberrant DNA structures. One such structure is DNA G-quadruplexes (G4s), which have broad regulatory functions. Although putative G4-forming sequences have been previously implicated in BS, it remains unclear what (dys)regulatory role, if any, endogenous G4 structures may play in BS. Here, we profiled chromatin accessibility and gene expression via ATAC-seq and RNA-seq and mapped endogenous G4 via ChIP-seq in wild-type (WT) vs. BS cell lines. We observed that in BS, differential G4 formation positively correlated with both chromatin accessibility and gene expression. Stabilizing G4 in WT cells with pyridostatin partially phenocopied BS. Additionally, data from a BS family showed that regions with increased chromatin accessibility in BS individuals also were enriched for G4-forming sequences. Our data showed that G4 formation is associated with higher chromatin accessibility and gene expression; likely in BS, unresolved G4 increases focal chromatin accessibility, thereby upregulating gene expression. In summary, our results revealed a central role of G4 in the molecular etiology of BS and provide a new perspective on BLM's regulatory function through G4s.

## Introduction

In humans, Bloom Syndrome (BS, OMIM 210900) is an autosomal recessive cancer-predisposition disorder caused by loss-of-function mutations in the *BLM* gene, which encodes a 3' to 5' DNA helicase in the evolutionarily conserved RecQ helicase family [1–4]. BLM helicase unwinds various abnormal DNA structures formed during DNA replication, repair, and recombination and thus has a crucial role in maintaining genome integrity [5–7].

Loss of BLM helicase function has detrimental effects at both cellular and organismal levels. BS individuals manifest multifaceted clinical phenotypes [4]. Besides the prominent dwarfism from pre- and postnatal developmental delay [4,8], BS individuals are prone to develop cancer early and have a strikingly shorter lifespan, on average

of 26 years. Molecularly, BLM-deficient cells exhibit widespread molecular alterations. The characteristic feature and diagnostic criterion of Bloom syndrome (BS) is excessive sister chromatid exchange events (SCEs), which occur at a rate seven to ten times higher than in normal cells [9,10]. This echoes the indispensable function of BLM helicase in resolving double Holliday junctions without generating cross-over DNA products and thus suppressing illegitimate genetic recombination events [11]. This latter feature of elevated recombination during mitosis has been leveraged to study and compare gene functions [12,13]. In addition, BLM-deficient cells also display features underlying genome instability, impaired DNA repair, and elevated frequency of loss of heterozygosity, all of which increase the risk of developing cancer [14–16]. Additionally, BS cells have a slower cell cycle, altered gene expression profile, and increased epigenetic age in terms of their DNA methylation patterns [17–20]. Nonetheless, it is generally accepted that the loss of BLM's function in maintaining genome integrity is a central cause of BS.

One of the substrates of BLM helicase is G-quadruplexes (G4s) [21], which are four-stranded secondary DNA structures formed by G-rich DNA sequences (also known as G4 motifs). Rather than G-C base pairing, four guanine pair with each other through a cyclic Hoogsten hydrogen bonding and form a square planar structure, G-quartet. Two or three G-quartets subsequently stack on top of each other to form a G4. Bioinformatic tools can predict and detect putative G4-forming sequences in DNA sequences, known accordingly as G4 motifs [22,23]. With canonical G4 motifs ($G_{3+}$ $N_{1-7}$ $G_{3+}$ $N_{1-7}$ $G_{3+}$ $N_{1-7}$ $G_{3+}$), depending on the underlying algorithms, 400,000 to over a million G4 motifs *in silico* can be detected in the human genome [24,25]. Another approach, G4-seq, which takes advantage of G4-induced polymerase stalling, identified over 500,000 G4-forming sequences (herein referred to as G4-seq hits) *in vitro* in purified genomic DNA [26]. In contrast, G4 ChIP-seq or G4 CUT&Tag using a specific antibody BG4 against G4 structures in cells only identifies 10,000 to 20,000 endogenous formed G4s in cells [27–31].

G4s are highly enriched in telomeres, enhancers, and promoters and they have been proposed or found to be able to regulate various processes such as DNA replication, transcription, and DNA-protein interaction [22,29,32–34]. For instance, the formation of G4 could impede the progression of the DNA replication fork or machinery, which causes

DNA damage and initiates DNA repair pathways [35,36]. G4 structures have also been implicated in gene regulation. The formation of G4 could impair the loading of RNA polymerase II, modulate the formation of R-loops, or stabilize the transcription bubble, thereby suppressing or enhancing transcription [37,38]. The expression of many oncogenes, such as *c-MYC*, *c-KIT* and *KRAS* is influenced by the G4 forming status in their promoter and G4s are therefore considered a therapeutic target in cancer [39–44]. Additionally, the formation of G4 may affect DNA-protein interaction [29,32,45]. For instance, G4 formation may recruit transcription factors or even hinder histones binding to DNA, which could alter the local chromatin accessibility.

Interestingly, recent studies have hinted at the relevance of G4 motifs in the molecular phenotypes of BS [9,19,46]. It has been reported that differentially expressed genes in BS were correlated with the presence of G4 motifs [19,46]. Moreover, van Wietmarschen et al. demonstrated that SCEs, the signature in BS, are enriched at G4 motifs in BS [9]. These possible connections notwithstanding, the previous analyses have been limited to G4 motifs and not endogenous G4 structures. Crucially, the presence of G4 motifs alone is a poor predictor of G4 formation in living cells. Therefore, it remains unknown how the loss-of-function of BLM helicase affects the endogenous G4 formation and its significance in the molecular phenotypes in BS. Considering the regulatory function of G4s, we thus set out to investigate the possible link between BS and G4 by firstly, characterizing the molecular changes in BS cells, and secondly, determining if these changes are associated with endogenous G4 formation, and the changes in G4s.

In this study, we assayed the molecular changes via ATAC-seq and RNA-seq in cell lines derived from healthy donors and BS individuals. Importantly, rather than using G4 motifs or G4-seq hits as a proxy, we performed G4 ChIP-seq to directly map endogenous G4s in healthy and BS cells. We first compared molecular profiles in BS vs. WT cells and observed that BS cells exhibited distinguished G4 profiles. Notably, the differential G4 formation in promoters positively correlated with the changes in both chromatin accessibility and gene expression. Further, we explored the causal role of G4 and molecular changes in BS. Treating WT cells with pyridostatin (PDS), a small molecule that stabilizes G4 structures, to mimic the defective ability to resolve G4s in BS, partially recapitulated the observed chromatin accessibility and gene expression changes in BS. Moreover, ATAC-seq data from a BS family showed that more

accessible chromatin regions in BS individuals exhibited stronger enrichment for G4-seq hits. We also observed that the presence of G4 was associated with higher chromatin accessibility and gene expression. Possibly there is a feedback mechanism between G4 and ATAC-seq, and differential G4 formation regulated gene expression via regulating chromatin accessibility. These changes were enriched in molecular pathways relevant to the clinical manifestations of BS. Our results demonstrate that G4 formation serves as an additional layer of regulation in BS and it is associated with higher chromatin accessibility and higher gene expression. Therefore, we hypothesize a novel model of G4 mediating molecular changes in BS. Our study is the first study to map endogenous G4 in BS and provides direct evidence that G4 could emerge as a central factor in the molecular etiology of BS and potentially serve as a therapeutic target.

## Results

### Bloom Syndrome cells exhibit distinct molecular profiles

To characterize the molecular changes in BS, we collected two pairs of lymphoblastoid (LCL) and fibroblast cell lines (Fib) derived from sex-matched and roughly age-matched affected BS individuals and healthy donors (wildtype, or WT). We assayed the epigenetic changes via ATAC-seq and profiled transcriptomic changes via RNA-seq in BS vs. WT cells. Importantly, instead of inferring the G4 formation by *in silico* predicted G4 motifs or *in vitro* validated G4-seq hits, we utilized G4 ChIP-seq [27] and an antibody specific to G4 structures to map the endogenous G4s in cells (**Fig. 1a**).

We first employed principal component analysis (PCA) to compare the global molecular profiles in WT lymphoblastoid (LCL-WT), BS lymphoblastoid (LCL-BS), WT fibroblast (Fib-WT), and BS fibroblast (Fib-BS) cell lines. PCA showed consistent patterns across all three types of data: the major source of variation, PC1, accounted for over 70% of the variance and predominantly separated Fib and LCL samples, suggesting tissue-specific molecular profiles (**Suppl. Fig. 1a**); PC2 and PC3 separated WT samples from the BS samples in LCL and Fib cell lines (**Fig. 1b**), respectively, implying distinct chromatin accessibility, gene expression and G4 profiles in BS and cell-type specific impacts of BS.

We then carried out differential analyses to characterize the molecular alterations in BS. Firstly, in ATAC-seq data, we identified 101,787, 115,027, 125,352, and 114,427 open chromatin regions in LCL-WT, LCL-BS, Fib-WT, and Fib-BS cells, respectively, and detected the differential signals with *DESeq2* [47]. Of note, we applied a local correction to the differential analysis of Fib samples which showed copy number differences (see Methods) [48]. In the LCL-BS vs. LCL-WT comparison, we identified 72,580 (57%) significantly differentially accessible (DA) chromatin regions (adjusted p-value (p-adj) < 0.05), of which 39,920 and 32,660 displayed increased (BS > WT) and decreased (BS < WT) chromatin accessibility (**Suppl. Fig 1a, left**). In the Fib-BS vs. Fib-WT comparison, we employed copy number normalization [48] in the differential analysis and identified 36,225 (25%) and 42,564 (35%) peaks with significantly increased and decreased chromatin accessibility, respectively (**Suppl. Fig. 1a, right)**. Notably, DA peaks identified in LCL-BS and Fib-BS showed a weak anti-correlation (Pearson correlation, r = -0.11, $P < 0.001$; **Suppl. Fig. 1b**) despite about 8,000 shared DA peaks, implying the relatively stronger cell-type specific impacts on the chromatin states of BS.

Secondly, we characterized gene expression changes in BS for each cell line. We utilized transcripts per million (TPM) as normalized gene expression values and defined expressed genes as those with TPM-transformed z-scores ≥ -3 [49,50]. Excluding genes not expressed in WT or BS cells. Via *DESeq2* we identified 6,703 significantly differentially expressed (DE) genes (p-adj < 0.05; 3,627 up-regulated and 3,076 down-regulated) in LCL and 11,140 significantly DE genes (p-adj < 0.05; 5347 up-regulated and 5793 down-regulated) in Fib (**Suppl. Fig. 1c**). Consistent with the cell-specific changes in BS in the PCA, the DE gene sets in LCL and Fib were distinct and displayed weak correlation (Pearson correlation, $r = -0.03$, $P < 0.001$; **Suppl. Fig. 1d**). We further asked if we can recover consistent molecular functions among DE genes with Gene ontology (GO) analysis using gene lists ranked by the fold changes. In general, down-regulated genes were associated with more cell type-specific pathways. In LCL, we found enriched pathways were mostly related to immune functions such as the “activation of the immune responses” and “immunoglobulin production”, implying potentially decreased immune functions in BS, which aligns with the observed immunodeficiency in BS individuals (**Suppl. Fig. 1f**) [4]. In Fib, they were mostly related

to cell skeleton organization and cell morphology-related structure development, such as muscle structure development. In both cell types, most of the enriched GO terms from the upregulated genes are associated with ion channel activities and cell transport (**Suppl. Fig. 1f**). Notably, both cell types showed enrichment in the hallmark gene set of MYC targets, which is closely related to tumorigenesis (**Suppl. Fig. 1g**). *MYC* is a well-known gene whose expression level is highly dependent on the formation of G4 in its promoter [39,40]. Furthermore, to determine the universal impacts of BS, we analyzed the enriched molecular pathways of shared up-regulated or down-regulated genes in different cell types. We found that the up-regulated genes were enriched for DNA-transcription factor activity, cell junction, cell adhesion, and signaling transduction (**Fig.1c, left**). In contrast, the down-regulated genes were associated with translation, ribosome metabolism, and ncRNA processing (**Fig.1c, right**). Of note, BLM has been reported to bind to facilitate RNA polymerase I-mediated ribosomal RNA transcription through unwinding G4s [51,52].

Thirdly, using *MACS2* [53] to call G4 peaks and *IDR* [54] with the criteria of irreproducible rate (IDR) < 0.05 to filter the reproducible peaks across biological replicates, we identified 10,382, 15,204, 17,125, and 13,415 high-confidence endogenous G4s in LCL-WT, LCL-BS, Fib-WT and Fib-BS cells, respectively. Only 4,977 G4 sites (44.8%, 35.8%, 29.4% and 38.4% in LCL-WT, LCL-BS, Fib-WT, Fib-BS, respectively) were shared across all samples, showcasing that cell- and genotypes have a strong effect on the set of endogenous G4 (**Fig.1e, Suppl. Fig. 2a**). About 80% of detected G4 peaks overlapped with G4-seq hits (G4-forming sites validated in the purified genomic DNA *in vitro* via G4-induced polymerase stalling) (**Suppl. Fig. 2c**) [35]. These G4 peaks were enriched DNA motifs with triple and/or quadruple Guanine-repeats, featuring G4-forming sequences (**Suppl. Fig. 2b**).

Endogenous G4 is frequently observed in the promoter of highly expressed housekeeping genes and some oncogenes [27,29,34,40,41,43]. In our data, we detected G4 peaks in the housekeeping genes such as *GAPDH* and *H3*, in oncogenes such as *KRAS*, *MYC*, and *KIT*, and less reported but a more LCL-specific G4-forming gene, *NFKB1* (**Fig.1d**) and these G4 peaks overlapped with open-chromatin signatures (**Fig.1d**). Given the nature of G-quadruplex formed by single-stranded DNA, G4s are expected to form in open chromatin regions. Overall, about 70% ~ 80% of G4 peaks

overlapped with ATAC-seq peaks (**Suppl. Fig. 2d**), which was reported in previous studies [29,30]. In terms of genetic elements, G4 signals are highly abundant in the gene body, particularly near the TSS. Approximately 80% of G4 peaks resided in the promoters (defined as TSS ± 3kb) with more than 50% directly overlapping a TSS (**Suppl. Fig. 2e, 2f**). Moreover, G4 peaks exhibited a significant enrichment in the 5' UTR, TSS, first exon, first intron, first exon-intron (1st-Ex-Int) junction, and 3' UTR compared to randomly shuffled G4 peaks regions of the same length (permutation, n = 1000, $P < 0.001$ for all features; see Methods) (**Suppl. Fig. 2g**). Notably, G4s displayed highly similar signal profiles to ATAC-seq in promoter regions with both signals peaking slightly 5' to the TSS (**Fig. 1f**). These observations that G4s predominantly locate in functional genetic elements strongly suggest that they could serve as an additional layer in the gene regulatory network.

To address whether the balance of G4 formation and resolution is disrupted in BS and the significance of G4s in BS, we characterized the differential G4-forming sites (DF G4 sites) in BS cells with *DiffBind* [56,57]. In LCL-BS vs. LCL-WT, we identified 2,436 (13.79%) significantly DF G4 sites (FDR < 0.05), 1,401, and 1,035 with increased and decreased G4 formation, respectively (**Suppl. Fig. 1e, left**). Similarly for ATAC-seq, we applied copy number normalization [48] in Fib samples and detected 2,022 (10.07%) and 1,764 (8.49%) sites with increased and decreased G4 formation (**Suppl. Fig. 1e, right**). Compared to the distribution of the G4 peaks in different genetic elements, the DF G4 sites are not particularly enriched for certain genetic elements (**Suppl. Fig. 2e**)

To summarize, BB cells harbor disrupted molecular profiles across all data modalities. Particularly, BS cells displayed a distinct G4-formation landscape. Additionally, BS showed different impacts on different cell types.

## No particular enrichment of endogenous G-quadruplexes at the first exon-intron junction of differentially expressed genes in Bloom Syndrome cells

Previous work showed contradictory results on whether there is an enrichment of G4-forming sequences at the 1st-Ex-Int junctions of DE genes in BS [9,46,58]. Crucially, the presence of putative G4-forming sequences does not reliably predict the endogenous formation of G4 *in vivo*. Having obtained G4 ChIP-seq peaks allows us to address this question directly.

We first assessed the background enrichment of endogenous G4 peaks at the 1st-Ex-Int junction of expressed genes by comparing the number of observed and expected overlaps. The expected overlap was determined through permutation: randomly shuffling G4 peaks in the genome and counting overlaps with 1st-Ex-Int junction sites of expressed genes (LCL: $P$ <0.001; Fib: $P$ < 0.001) (**Suppl. Fig. 3a, 3e**). Compared to the background enrichment level, DE genes in LCL-BS vs. LCL-WT did not particularly show a significantly stronger enrichment (Permutation, n=1000; $P$ = 0.15) (**Suppl. Fig. 3b)**. Nonetheless, we did observe a significant enrichment of G4 at the 1st-Ex-Int junction of the DE genes in Fib-BS vs. Fib-WT (Permutation, n=1000; $P$ < 0.001) (**Suppl. Fig. 3f)**. Since most of the G4 peaks intersected with TSS, we also assessed and compared the enrichment of G4 at TSS of DE genes and expressed genes. Similarly, G4s were enriched at the TSS of expressed genes in both cell types (Permutation, n=1000; $P$ < 0.001). However, compared to this background enrichment, DE genes showed no particular enrichment in LCL (Permutation, n=1000; $P$ = 0.20) albeit in Fib (Permutation, n=1000; $P$ < 0.001) (**Suppl. Fig. 3d, 3h**). To summarize, instead of inferring the significance of G4 in the DE genes in BS by G4 motifs and G4-seq hits, using endogenous G4-peaks, we did not observe a consistent enrichment at the 1st-Ex-Int junction of DE genes in BS.

**G4 formation changes positively correlate with chromatin accessibility changes in Bloom Syndrome cells**

Next, we looked at how gene regulatory activity may be linked across different modalities We first looked at the DA peaks and DE genes. To link ATAC-seq peaks to their potential target genes, we assigned the closest ATAC-seq peak to the TSS in the promoter of a gene as the main putative regulatory element. In both cell types, we observed a strong positive correlation between DA peaks in promoters and DE genes (Pearson correlation; LCL, $r$ = 0.58, $P < 2.2\times10^{-16}$; Fib, $r$ = 0.53, $P < 2.2\times10^{-16}$) (**Fig. 2a, Suppl. Fig. 2b)**.

To further explore the relationship between differential chromatin accessibility and G4 formation in BS cells, we intersected the ATAC-seq peaks and G4 sites for each cell line with a minimum overlap of 150 bp. With this criterion, we identified in total 12,680 and 16,345 ATAC/ G4 peaks overlapping peaks in LCL and Fib, respectively. Across

this set, we found significant correlation in the directional changes in chromatin accessibility and G4 formation (Pearson correlation; LCL, $r = 0.58$, $P < 2.2\times10^{-16}$; Fib, $r = 0.59$, $P < 2.2\times10^{-16}$) (**Fig. 2b, Suppl. Fig. 4b**). Specifically, as the G4 ChIP-seq signal gradually transitioned from decreased to increased formation in BS, these ATAC-seq peaks showed coherently decreased to increased chromatin accessibility in BS. When there were increased G4 ChIP-seq signals ($log_2FC > 0$) in BS, these regions also showed overall increased chromatin accessibility ($log_2FC > 0$) in BS and vice versa (**Fig. 2c, right; Suppl. Fig. 4c, right)**. In regions where G4 ChIP-seq signals noticeably dropped in BS, the ATAC-seq predominantly exhibited decreased chromatin accessibility, with only a few displaying increased accessibility (**Fig. 2c, left; Suppl. Fig.4c, left)**. Conversely, in BS, regions with remarkably higher G4 ChIP-seq signals are predominantly more accessible in BS and rarely less accessible (**Fig. 2c, left; Suppl. Fig.4c, left)**. The positive relationship implied a tight link between G4 formation activities and chromatin accessibility in Bloom Syndrome. Nonetheless, it is difficult to distinguish which signal change is primary to the other and either way is possible.

**G4 formation changes positively correlated with gene expression changes in Bloom Syndrome cells**

Given the proposed role of G4 in regulating gene expression, we next asked whether G4 formation activities correlate with differential gene expression. To do so, we linked G4 peaks to a target gene if the G4 peak resides with the gene's promoter. This way, we identified in total 9272 and 11657 G4-genes in LCL and Fib, respectively.

Interestingly, we found that gene expression changes showed a significant positive correlation with G4 formation changes (**Fig. 2d**). Changes in gene expression significantly positively correlated with changes in G4 formation (Pearson correlation; LCL, $r = 0.28$, $P < 2.2\times10^{-16}$; Fib, $r = 0.37$, $P < 2.2\times10^{-16}$) (**Fig. 2d, Suppl. Fig. 4d**). Similar to what we have observed in the differential ATAC-seq signals, as G4 ChIP-seq signal transitioned from decreased to increased formation in BS, these genes with G4 in their promoter changed consistently from decreased to increased expression in BS. Upregulated genes mostly have increased G4 formation in the promoter and they showed overall increased gene expressions in BS (**Fig. 2e**). Conversely, less G4-

forming sites mainly coincided with downregulated genes and overall lower gene expression (**Fig. 2e**).

Taken together, gene expression changes in Bloom Syndrome showed a strong positive correlation with changes in G4 peaks in their promoters. Increased G4 formation in promoters was linked to higher gene expression, suggesting that G-quadruplexes could represent an additional layer of transcriptional regulation in BS. This result goes beyond previous work that hinted at the association between G4-forming sequences and BLM-regulated genes [19,46,58].

**Endogenous G-quadruplexes are enriched in the sister chromatid exchange events in Bloom Syndrome cells**

Previously, van Wietmarschen et al. applied single-cell DNA template strand sequencing (Strand-seq) to map the genomic locations of SCEs, the molecular signature of BS, in both BS and WT cell lines at kilobase resolution [9]. They showed while SCEs in the WT do not appear to have a hotspot, in BS cells they are enriched for G4 motifs. Expanding on this observation, we obtained the SCE location mapped in the same BS-LCL cell line used in this study and assessed if they are also enriched at the endogenous G4 forming sites from the same cell line. Those SCE events overlap with G4 peaks significantly more often than expected by chance in the permutational analysis (permutation, n=1000; all SCEs, z-test, $P = 0.014$) (**Fig. 2f; Suppl. Table 1**). In addition, the SCEs were also enriched in open chromatin regions (permutation, n=1000; all SCEs, z-test, $P < 0.001$) (**Suppl. Fig. 4f; Suppl. Table 1**). We showed that SCEs were enriched at the endogenous G4 sites and provided direct evidence to the model that in BS, G4s could persist and accumulate in cells, which stall DNA polymerase and initiate HR DNA repair. Subsequently, due to the lack of BLM to suppress cross-over products in the last steps of HR repair, this eventually resulted in elevated SCE frequency [59].

**PDS treatment partially recapitulates the molecular changes in Bloom Syndrome cells**

Thus far, we demonstrated the widespread changes in the molecular profiles in BS were inter-connected. Both the changes in chromatin accessibility and gene expression significantly correlated with G4 formation activities, which addressed the association between G4 and BS. Next, we sought out to investigate the causal role of

G4 in those observed molecular changes in BS by employing pyridostatin (PDS), a commonly used G4-stabilizing small molecule [60]. Moreover, it has been validated via *in vitro* biochemical assays that the application of PDS makes it more difficult for BLM to unwind G4 structures [61]. We therefore applied PDS to WT cell lines to mimic the defective ability to resolve G4 structures in BS, and profiled the open chromatin landscape and gene expression profiles in treated vs. control cells for both cell types (**Fig. 3a**).

To assess the overall impacts of Bloom Syndrome and PDS treatment, we performed PCA to compare the global ATAC-seq profiles across all samples, including both PDS and BS contrasts. PCA revealed distinct chromatin accessibility profiles among samples of different tissue backgrounds, with PC1 explaining approximately 80% of the variance (**Fig. 3b, left**). Notably, WT samples, control samples, PDS-treated samples, and BS samples formed a spectrum on PC4 (**Fig. 3b, left**). This axis highlighted the common impacts between Bloom Syndrome and PDS treatment on chromatin accessibility, irrespective of the cell type. PC2 (7% of the total variation) and PC3 (6% of the total variation) separated Bloom Syndrome samples from the remaining samples in LCL and fibroblasts, respectively, indicating more cell-specific effects (**Suppl. Fig. 5a**). Further, we carried out PCA on gene expression profiles across all the samples and observed similar results as in ATAC-seq. The primary source of variation was attributed to different cell types (PC1) (**Fig. 3b, right**). Notably, PC4 aligned the samples along the same WT-control-BS-PDS continuum, suggesting common impacts between BS and PDS on the transcriptome as well (**Fig. 3b, right**).

Next, we characterized the differential signals. In LCL, PDS treatment led to 33,734 (27%) regions with significantly changed chromatin accessibility, 17,154 and 16,580 peaks with increased and decreased signals, respectively (**Suppl. Fig. 5b, left**). In Fib, we detected 42,301 (32%) DA peaks, of which 20,210 were more accessible and 22,091 were less (**Suppl. Fig. 5b, right**).

To assess whether PDS treatment elicits similar changes in chromatin accessibility as observed in BS for each cell type, we compared differential signals caused by BS and upon PDS treatment and observed a significant positive correlation between them (Pearson correlation; LCL, $r = 0.29$, $P < 2.2\times10^{-16}$; Fib, $r = 0.44$, $P < 2.2\times10^{-16}$). For

convenience, here significantly increased and decreased signals in BS vs. WT were referred to as BS+ and BS-, respectively, and similarly for PDS vs. control. In LCL, 4,446 regions exhibited consistently increased chromatin accessibility in both BS and PDS-treated conditions (BS+ PDS+), while 7,608 regions exhibited decreased signals (BS- PDS-) (**Fig. 3c, top-left**). Similarly, in Fib, 9,060 and 15,036 regions showed coherent changes in both conditions (**Fig. 3c, top-right**). In both cell types, the observed numbers of BS- PDS- peaks were significantly higher than expected, with approximately two-fold enrichment (hypergeometric test; Fib, $P_{(> observed)} = 1.6\times10^{-6}$; LCL, $P_{(> observed)} = 0.003$) (**Fig. 3c, bottom**). The observed overlap of the BS+ PDS+ signals is also significantly higher in fibroblasts albeit not significant in LCL (**Fig. 3c**; hypergeometric test; Fib, $P_{(> observed)} < 1\times10^{-16}$; LCL, $P_{(> observed)} = 1.00$). In contrast, there were less than expected BS- PDS+ peaks in both cell types albeit only a significant depletion of BS+ PDS- peaks in Fib (**Fig. 3c;** hypergeometric test; $P_{(< observed)}$ in **Suppl. Table 2**).

For RNA-seq data we identified 6,258 (39%) DE genes in LCL (2,930 upregulated and 3,328 downregulated) and 4,828 DE genes (2,363 upregulated and 2,465 downregulated) in Fib (**Suppl. Fig. 5d)**, both of which strongly correlated with the DE genes in the corresponding BS cell line (Pearson correlation; LCL, r = 0.29, $P < 2.2\times10^{-16}$; Fib, r = 0.44, $P < 2.2\times10^{-16}$). Similar to DA peaks, the observed number of BS+ PDS+ and BS- PDS- DE genes was significantly higher than expected by chance in both cell types, displaying a 1.2 to 2-fold enrichment (**Fig. 3d, bottom;** hypergeometric test; $P_{(> observed)}$ in **Suppl. Table 3**). Conversely, there was a depletion of BS- PDS+ and BS+ PDS- genes (**Fig. 3d, bottom;** hypergeometric test; $P_{(< observed)}$ in **Suppl. Table 2**).

To summarize, in both ATAC-seq and RNA-seq data, assessing the global profiles by PCA, samples formed the WT-control-BS-PDS axis in the PC4 regardless of cell types, suggesting common impacts of BS and PDS treatment on the molecular profiles. Further comparing the differential signals, G4 stabilization via PDS could partially recapitulate the molecular changes in BS. The concordant changes in both global and differential signals suggest potentially shared regulatory mechanisms between G4 stabilization and BS and a causal role of G4. Moreover, chromatin accessibility changes in response to PDS treatment imply a possible feedback loop between G4

structure and nucleosome positioning. In contrast to cell type-specific effects of BS (**Suppl. Fig. 5a**), we noticed that PDS treatment elicited more universal but less extensive responses on both chromatin accessibility and gene expression (**Suppl. Fig. 5c, 5e**).

## Differentially accessible chromatin regions in Bloom Syndrome individuals are associated with the prevalence of G4-seq hits

To validate our results from WT and BS cell lines, we further analyzed lymphocytes isolated from a Bloom Syndrome family with the most prominent BS-causing mutation, *BLM* c.2207- 2212delATCTGAinsTAGATTC Ashkenazi BS mutation (*blm*$^{Ash}$) (**Fig. 4a**; **Suppl. Table 3**) [62]. Due to the limited material, we were only able to assay the open chromatin landscape via ATAC-seq. Further data from the carrier mother and carrier daughter were excluded due to low data quality. In total, we identified 60,913 peaks from non-BS and BS individuals and 982 differentially accessible chromatin regions, 60 with increased signal and 922 with decreased signals (**Suppl. Fig. 6**). To determine if the differentially accessible chromatin regions are associated with G4 forming sequences, we intersected open chromatin regions with *in vitro* G4-seq hits due to the lack of corresponding endogenous G4 data for these samples. We examined the enrichment of G4-seq hits in ATAC-seq peaks by comparing the observed overlap to the expected overlap estimated by permutation, in which we counted the number of G4-seq hits in open chromatin regions randomly shuffled in the genome. Overall, open chromatin regions were enriched for G4-seq hits (permutation, n=1000; observed overlap: 38514; expected overlap: 8290 ± 108.8; z-test, $P < 0.001$; fold enrichment = 4.65). Interestingly, the more accessible chromatin regions in BS individual (permutation, n=1000; observed overlap: 24; expected overlap: 10.5 ± 3.81; z-test, $P < 0.001$; fold enrichment = 2.29) exhibited higher enrichment for G4-seq hits compared to those less accessible chromatin regions (permutation, n=1000; observed overlap: 85; expected overlap: 65.3 ± 8.55; z-test, $P < 0.01$; fold enrichment = 1.30) (**Fig. 4b**).

In terms of molecular functions, these differentially accessible chromatin regions were enriched in genes related to TNF-α signaling via NF-κB, responses to UV, apical junction, and epithelial-mesenchymal transition (**Fig. 4c**). This observation aligns with the fact that BS individuals are hypersensitive to UV light causing the rashes on their faces [4]. Moreover, G-quadruplexes have been implicated in the regulation of epithelial-

mesenchymal transition [63]. Consistent with our findings in cell lines, data from Bloom Syndrome individuals suggest that G4 is prevalent in differentially accessible chromatin regions which were associated with functions relevant to the clinical symptoms in BS, further highlighting the significance of G4 in the molecular etiology of BS.

**Molecular hints of increased G4 formation enhancing chromatin accessibility and gene expression**

So far, we have addressed the association and partial causal relationship between G4 and BS in cell lines and further highlighted the significance of G4-forming sequences in data from a Bloom Syndrome family. Next, we sought to disentangle possible molecular mechanisms of how the formation and/or differential formation of G4 feature in the molecular changes in BS. For clarity, we report the results from LCL-WT, but we stress that we found similar effects in the other cell types as well.

To determine if G4 formation may affect chromatin accessibility, the ATAC-seq peaks were stratified into two groups by whether they overlap with a G4 peak or not. We observed that ATAC-seq peaks overlapping G4 peaks are wider (median size: 988bp vs. 347bp; Wilcoxon test, $P < 2\times10^{-16}$) and exhibited overall higher chromatin accessibility (median FPKM: 31.1 vs. 5.96; Wilcoxon test, $P < 2\times10^{-16}$; **Fig. 5a**). Additionally, we assessed the density of open chromatin regions by counting the number of ATAC-seq peaks in the flanking 10kb region of each ATAC-seq peak (**Suppl. Fig. 7a**). G4 peaks resided in regions with denser ATAC-seq peaks (Wilcoxon test, $P < 2\times10^{-16}$; **Fig. 5b**), implying G4 formation may not only influence the overlapping ATAC-seq peaks but also potentially proximal peaks.

Given the previously observed positive correlation between DF G4 sites and DA peaks and the "nucleating" effect above, we asked if their positive correlation extended to the proximal ATAC-seq peaks besides the overlapping ones. To do so, we assigned ranked numbers to the ATAC-seq peaks within 10kb from the center of the G4 peak by their relative position to the G4 peak: 0 for overlapping peaks, positive values for upstream adjacent peaks and negative values for downstream adjacent peaks (**Suppl. Fig. 7b**). We observed the strongest correlation between the $\log_2$FC of the G4 sites and the overlapping ATAC-seq peak and the correlation diminished sharply for G4-

proximal ATAC-seq peaks (**Fig. 5c, top; Suppl. Fig. 9a, top**). The same dramatic decrease in correlation was also observed when we evaluated the proximity of ATAC-seq peaks to the G4 peaks using the physical distance (**Fig. 5c, bottom; Suppl. Fig. 9a, bottom**).

The positive correlation between G4 changes and chromatin accessibility changes cannot directly address the hierarchy between G4 and open chromatin. Intuitively, an open chromatin state is the prerequisite for G4 formation. However, G4s were in regions with denser ATAC-seq peaks and their presence is associated with wider and more accessible ATAC-seq peaks, suggesting that the formation of G4 could potentially feedback to regulate chromatin accessibility. In particular, the potential modulation of G4 on chromatin accessibility was mostly restricted to the overlapping open chromatin regions.

To determine if G4 formation may regulate gene expression, we similarly stratified the expressed genes into those with G4 in their promoters vs. those without. Interestingly, genes with a G4 signature in their promoter exhibited higher gene expression levels (median TPM: 22.2 vs. 10.8; Wilcoxon test, $P < 2\times10^{-16}$) (**Fig. 5d**). Similar to the distance effect we observed above, we also noted that the average expression effect decreases with distance to the G4 peak (**Suppl. Fig. 7c**).

It has been reported that G4 formation on the transcribing or non-transcribing strand in the promoter could have a different influence on gene expression [37]. To dissect this effect on transcription in our data, we intersected our G4 peaks with the publicly available G4-seq data [26] to assign the strandedness of G4 peaks, which by itself does not retain strand information. Specifically, the G4 structure is postulated to form on the Watson (or Crick) strand if the G4 peak only intersects G4-seq hits on Watson (or Crick) strand, (**Fig. 5e, left**). Otherwise, the strandedness cannot be inferred if the G4 peak does not overlap any G4-seq hit or overlap G4-seq hits from both strands (**Fig. 5e, left**). This way about 50% of all the G4 peaks can be assigned to a strand and their strandedness relative to the transcription direction of their target gene (**Fig. 5e, right; Suppl. Fig. 7d**). We found that genes with G4 regardless of their strandedness showed higher expression compared to genes without G4 (**Fig. 5f,** Wilcoxon test, $P < 2\times10^{-16}$). Notably, genes with G4 on the non-transcribing strand showed relatively

higher gene expression than genes with G4 on the transcribing strand (**Fig. 5f,** Wilcoxon test, $P = 0.0008$). Given this impact of G4 strandedness on gene expression, we asked whether the strandedness of G4 makes a difference in the positive correlation between DF G4 and DE genes. We observed a consistent positive correlation regardless of the strandedness of G4 (**Suppl. Fig. 7e**). In summary, the formation of G4 was associated with higher gene expression, particularly when G4 forms on the non-transcribing strand. This could explain that in BS vs. WT, genes with increased G4 formation mostly showed increased gene expression. In contrast, the strandedness of G4 showed little influence on chromatin accessibility (**Suppl. Fig. 7f**).

Next, we delved into the molecular mechanism of G4's regulatory function from the perspective of differential signals. Given the three-way correlation between ATAC-seq, G4 ChIP-seq and RNA-seq, we aimed to explore the interplay between G4 changes and chromatin accessibility changes on gene expression changes in the subset of genes that were expressed and contain both ATAC-seq and G4 peaks. We identified 9,194 and 11,580 such genes in LCL and Fib, respectively.

We first determined which signal changes or if both, differential ATAC-seq or differential G4 ChIP-seq influences gene expression. By comparing the strength of gene expression changes of genes stratified by the differential status of the ATAC-seq and G4 peaks, we found that genes with significant change in either modality showed increased expression (Wilcoxon test; *, $P < 2\times10^{-16}$) (**Fig. 5g, Suppl. Fig. 9b)**; but the greatest increase were observed in genes showing significant changes in both chromatin accessibility and G4, suggesting an additive, if not a synergistic interaction (Wilcoxon test; #, $P < 2\times10^{-16}$) (**Fig. 5g, Suppl. Fig. 9b).**

Given the positive correlation between DA peaks and DF G4 sites, we next investigated if there was any hierarchy between differential ATAC-seq and differential G4 ChIP-seq signals in regulating gene expression. To do so, we evaluated the correlation between gene expression changes and another data modality stratified by the differential status of the third modality. Interestingly, significant changes in G4-formation showed weak impacts on the positive correlation between differential chromatin accessibility and differential gene expression whereas the positive correlation between differential G4 ChIP-seq and differential gene expression largely

depended on significant changes in chromatin accessibility (**Fig. 5h; Suppl. Fig. 8; Suppl. Fig. 9c, 9d**). This suggests that altered G-quadruplexes potentially regulate gene expression mainly by modulating chromatin accessibility rather than the other way around and importantly, it also implies the reciprocal modulation between G4 and ATAC-seq.

To formally test if our data supports the feedback mechanism wherein G4 formation promotes further opening of chromatin, we built a multifactor ANOVA model to test for the interactive effect between ATAC-seq and G4 ChIP-seq (RNA = ATAC + G4 + ATACxG4; RNA being the gene expression change, ATAC and ChIP variables being their differential status). Our analysis revealed a significant interaction between chromatin accessibility and G4 formation (Type III ANOVA; F = 3.89, $P_{(>F)}$ = 0.004). Additionally, it confirmed that both differential ATAC-seq and differential G4 formation status imposed significant impacts on gene expression changes (Type III ANOVA; ChIP: F = 184.2, $P_{(>F)} < 2\times10^{-16}$; ATAC: F = 158.11, $P_{(>F)} < 2\times10^{-16}$).

To summarize, a significant change in G4 was associated with stronger changes in chromatin accessibility and gene expression. Changes in chromatin accessibility and G4 formation modulated transcription output cooperatively. Importantly, altered G4 formation in the promoter most likely affected gene expression by regulating chromatin accessibility.

## Discussion

Previous studies implicated the relevance of G4 in the molecular changes in Bloom Syndrome by demonstrating the association of the local presence of G4-forming sequences, transcriptomic changes, and SCEs in BS cell lines. However, crucially, neither *in silico* predicted G4 motifs in the human genome nor *in vitro* validated G4-seq hits in purified genomic DNA underlie the G4-forming status *in vivo*. Our study is the first one to map the endogenous G4s in BS and directly address the relationship between the local presence and change of G4 and BS. Rather than inferring G4 formation via harboring G4-forming sequences, we carried out G4 ChIP-seq to map genome-wide endogenous G4 sites and profiled the gene expression changes via

RNA-seq and epigenetic changes via ATAC-seq in WT and BS lymphoblastoid and fibroblast cell lines.

Firstly, we demonstrated the association between changes from each assay. We found a positive correlation between changes in endogenous G4 formation and molecular changes in BS. Increased G4 formation in BS was associated with increased chromatin accessibility and gene expression, and vice versa. Importantly, in the LCL-BS, SCE events, the signature of BS, were significantly enriched at endogenous G4 sites. Next, using PDS, we addressed the causation. We showed a consistent WT-control-BS-PDS spectrum on PC4, and enrichment of concordant changes between BLM deficiency and upon PDS treatment in both ATAC-seq and RNA-seq in both cell types. These suggest that G4 stabilization can partially phenocopy BS and some molecular changes in BS could be mediated through increased G4 formation. The significance of G4 in molecular changes in BS was further supported by data from the BS family. More accessible chromatin regions in BS individuals showed stronger enrichment for G4-seq hits. Our results for the first time provide direct evidence that one of the BLM's substrates, G4s, is a central factor in linking BLM deficiency to molecular changes and clinical manifestations in BS.

We also disentangled potential molecular mechanisms of how G4 could regulate chromatin accessibility and gene expression. G4s mostly resided in open chromatin regions and promoters. Notably, the presence of G4s was associated not only with wider, denser, and more accessible ATAC-seq peaks but also enhanced gene expression. By partitioning genes on the changes of G4 peak or ATAC-seq peaks, we observed that changes in chromatin accessibility and G4 formation acted cooperatively on the gene expression changes. Particularly, the correlation between gene expression changes and G4 formation changes was mainly observed when there was a significant change in the chromatin accessibility while a significant change in the G4 formation showed nuanced impacts on the correlation between the other two. Given the coherent changes between chromatin accessibility and G4 formation in BS, and G4 stabilization could alter chromatin accessibility, it is highly likely that in BS the increased (decreased) G4 formation directly increased (decreased) the focal chromatin accessibility, which subsequently led to enhanced or reduced gene expression.

G4 peaks were generally enriched for genetic elements of the expressed genes, such as the TSS, 5`UTR, the first exon, the first intron, and the 1st-Ex-Int junction. Previous studies reported contradictory results on whether G4 motifs are enriched at the 1st-Ex-Int junction of DE genes in BS [19,46,58]. Using the G4 peaks, compared to the background enrichment level of G4 in the expressed genes, the DE genes still showed significant enrichment of G4 peaks at the 1st-Ex-Int junction in Fib-BS but not in LCL-BS.

Although the G4 ChIP-seq assay does not capture strandedness information, we inferred which strand G4s form on by intersecting the G4 peaks and G4-seq hits. An *in vitro* study showed that G4 formation on the non-transcribing strand enhances transcription by promoting R-loop formation and elongation whereas G4 formation on the transcribing strand represses and even abolishes transcription [37]. However, we noted that genes with G4 on the non-transcribing strand displayed overall higher expression compared to those with G4 on the transcribing strand. Our observation does not contradict the earlier *in vitro* studies because G4 ChIP-seq is a bulk assay and the detected G4 peaks only represent the G4 formation status in a proportion of the cells. Though in these cells the gene expression is largely reduced compared to the genes without G4 formation, the averaged signal from the bulk would remain higher than those genes without G4 peaks, though it is lower than the genes with G4 on the non-transcribing strand. It would be advantageous to detect G4 at single-cell levels to directly address how the strandedness of G4 affects gene expression.

Notably, the molecular functions underlined by the differential signals in BS cells upon PDS treatment or in BS individuals were linked to G4-related processes and/or could help explain clinical manifestations of BS. For instance, the downregulated genes were mostly enriched for non-coding RNA, tRNA processing, and ribosome biogenesis. Multiple studies demonstrated that BLM directly binds rDNA repeats and regulates RNA polymerase I-mediated rRNA transcription [51,52,64]. Upregulated genes were highly enriched for cell-junction- and cell-adhesion-related pathways. Additionally, in both data from cell lines and the BS family, differential signals showed enrichment for epithelial-mesenchymal transition, which is characterized by the expression of certain cell adhesion molecules. It is critical for the development of many tissues and organs in the developing embryo, and numerous embryonic events such as gastrulation,

neural crest formation, heart valve formation, and myogenesis [65]. The dysregulation of this pathway could help explain the growth deficiency in BS individuals. Interestingly, this process is reported to be regulated by G-quadruplexes [63,66,67]. The epithelial-mesenchymal transition is also closely related to metastasis in cancer progression [65]. Another common GO hit is UV response, which can be explained by the function of BLM in DNA repair and matches the symptom of hypersensitivity to UV in BS individuals [4].

Based on our findings, we would like to propose the hypothesis that G4s mediate the molecular changes in BS and a novel corresponding model of G-quadruplexes in the molecular etiology of Bloom Syndrome (**Fig. 6**). During DNA replication, G4 formation could hinder the progression of the DNA replication machinery leading to polymerase stalling. With intact BLM, G4s are resolved to ensure the progression of DNA replication. In contrast, in BLM-deficient cells, unsolved G4s cause DNA damage, initiate the homologous repair (HR) pathway, and eventually lead to sister chromatid exchanges due to the lack of BLM to dissolve the double Holliday junction in the last steps of HR. Nucleosome-depleted DNA provides a niche for G4 to form. Our results suggest the reciprocal modulation between chromatin accessibility and G4 formation. The open chromatin state creates a favorable environment for G4 formation. With active BLM helicase, G4s in the open chromatin regions are unwound such that the chromatin switches to a closed state. Conversely, in BS unresolved G4 hinders the binding of histones to the focal DNA, causing the region to remain accessible and resulting in increased chromatin accessibly and wider and denser open chromatin regions. G4 may also hinder certain factors from recognizing and binding to focal DNA or recruit G4-binding proteins [32,39,45], influencing gene expression. G-quadruplexes are likely to form in the promoter of highly expressed genes. During transcription in BS, unresolved G4 formation in the promoter enhances gene expression, which is likely through increasing the focal chromatin accessibility. Notably, G4 on the non-transcribing strand boosts the gene expression to a larger extent compared to G4 forming on the transcribing strand.

G4 stabilization via PDS remarkably captured a substantial proportion (approximately 10% to 20%) of the changes observed in BS. Nonetheless, there are a few differences between PDS treatment and BS. Firstly, PDS treatment is short-term, whereas BS

imposes long-term effects on the cell lines. Secondly, G4 is one of BLM's substrates, and G4 stabilization via PDS mimics the universal loss of G4-resolving abilities, not specifically limited to those by BLM if there are helicase-specific subsets of G4. Moreover, the dynamics of G4 could differ in BS and upon PDS treatment. With PDS, the equilibrium of G4 strongly shifts toward G4 formation, and persistent G4 formation may render the focal chromatin rigid and impose even higher stress on molecular processes. While our results cannot fully address the causal role of G4, nonetheless, we could demonstrate that the relationship between G4 and molecular changes in BS is more than an association. To fully establish the causal role of G4, further studies are required, and more sensitive assays for mapping G4, such as CUT&Tag or even single-cell assays are required [70,71].

Although BLM helicase is not the only G4-unwinding enzyme, other helicases such as WRN, FANCJ, and DHX36 can also unwind DNA G-quadruplexes [72,73], and we acknowledge that there might be partial redundancy between them. However, other helicases cannot compensate for BLM [8] . Crucially, distinguished from others, BLM has a unique role in homologous recombination repair to suppress DNA recombination [74], and we observed that the excessive SCE events due to BS were enriched at G4 sites, further signifying the link between perturbed G4 dynamics and molecular phenotypes of BS. It would also be worth exploring the broader implications of G-quadruplexes in diseases caused by loss-of-function of other G4-unwinding helicases such as Werner Syndrome [5], to gain a deeper insight into the redundancy and specificity of different helicases for certain subsets of G4 structures.

To summarize, we provided direct evidence that G4 could emerge as a central factor in the molecular etiology of BS. Based on the association, partial causation, and validation from BS individuals, we propose a molecular model where upon the loss of function of BLM, defective G4-resolving abilities lead to G4 formation changes, which further leads to downstream molecular alterations in BS, contributing to the clinical phenotypes. Therefore, G4 may serve as a potential therapeutic target in BS. This study enriches the understanding of Bloom Syndrome at a molecular level and the regulatory function of G4 *in vivo*. Additionally, we revealed a more complex molecular function of BLM: it possesses broad regulatory potential through regulating G4s besides its well-established role as a helicase.

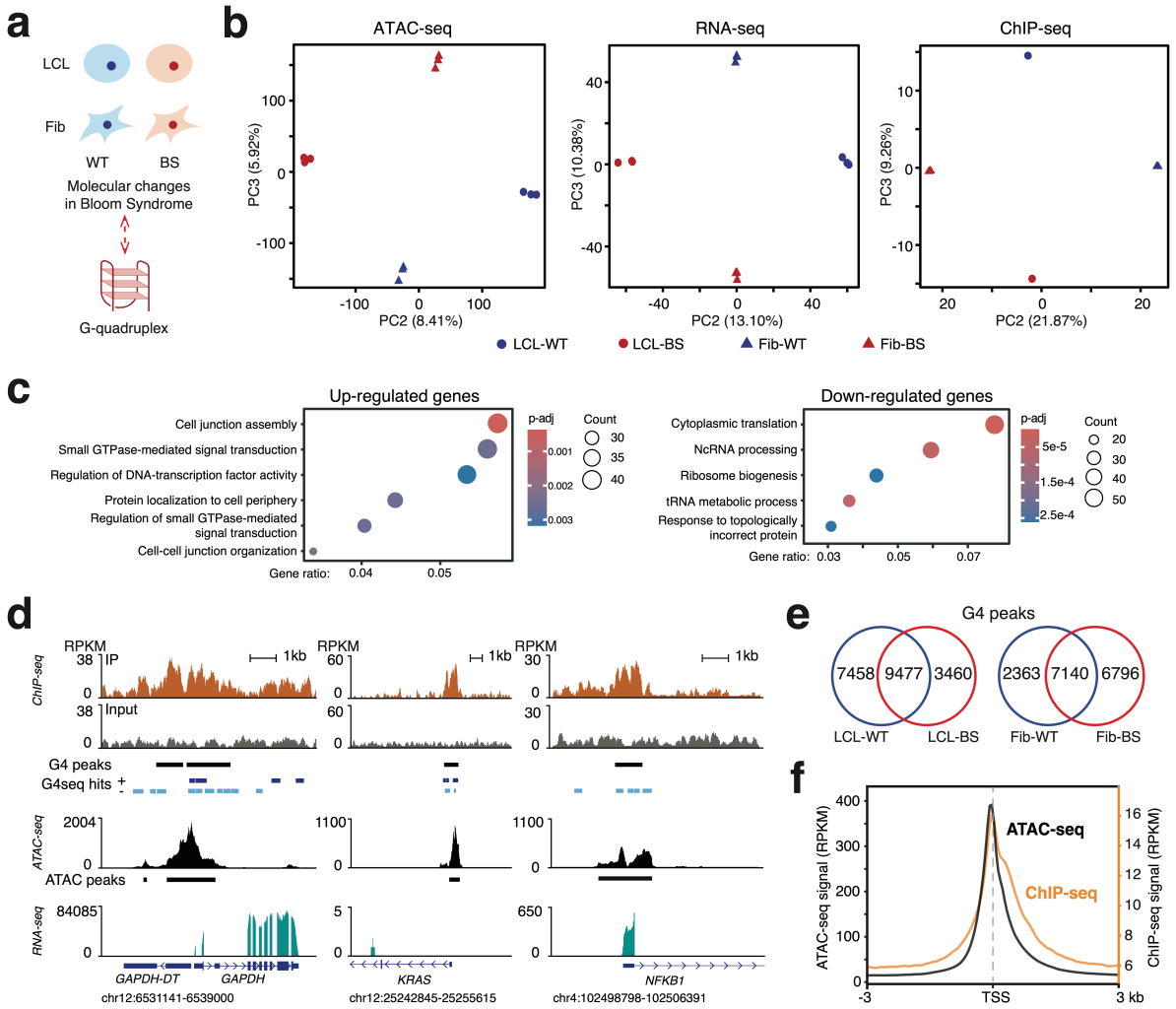



**Figure 1. Bloom Syndrome samples exhibited disrupted molecular profiles.** (**a**) Experimental design. LCL, lymphoblastoid cell lines; Fib, fibroblast cell lines; WT, wildtype, cell lines derived from healthy individuals; BS, Bloom Syndrome, cell lines derived from Bloom Syndrome individuals. (**b**) Principal component analysis of the genome-wide chromatin accessibility (left), gene expression profiles (middle), and endogenous G4 profile (right) in WT and BS samples. (**c**) Enriched molecular pathways of shared up- and down-regulated genes in LCL-BS and Fib-BS. (**d**) Genome browser screenshots showing the G4 ChIP-seq, ATAC-seq, and RNA-seq signals as well as G4-seq hits in example genes: a housekeeping gene *GAPDH*, an oncogene *KRAS,* and a less-known G4-forming gene *NFKB1*. (**e**) Number of high-confidence G4 peaks in different samples. (**f**) Averaged ATAC-seq and G4 ChIP-seq signal profiles in 6 kb regions flanking transcription start sites (TSS).

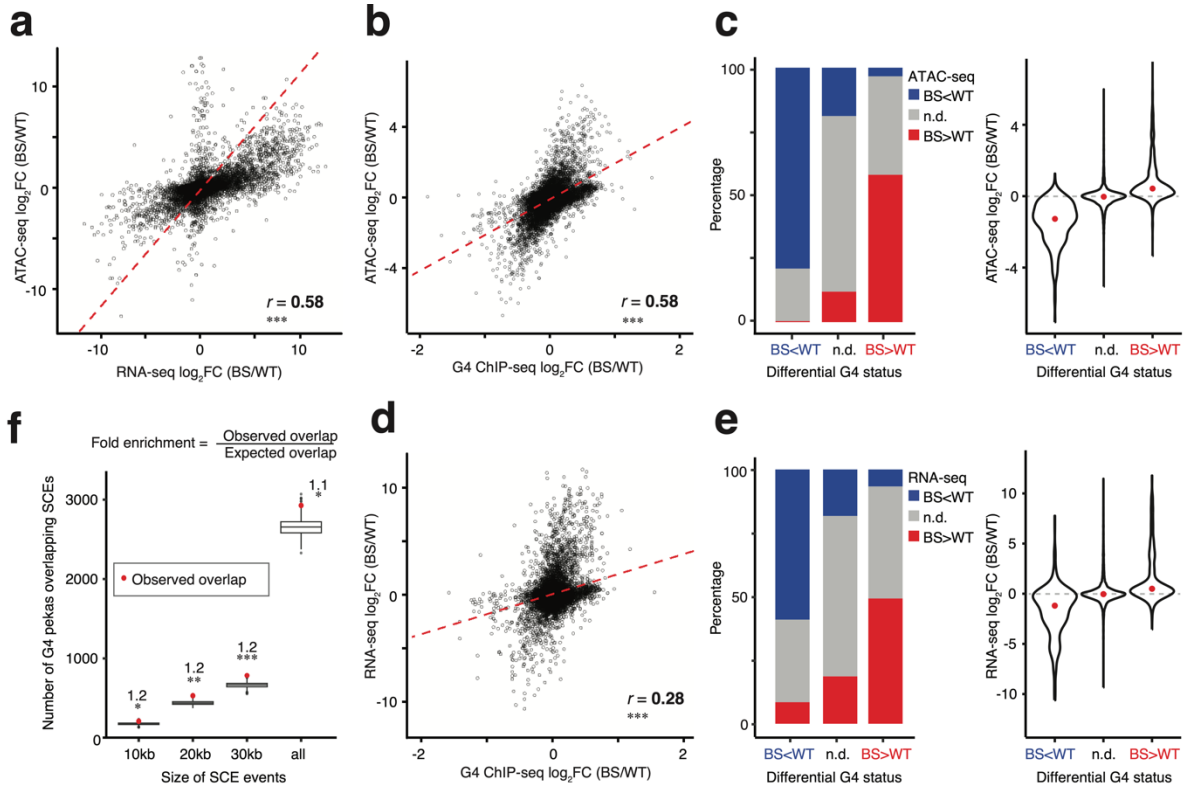



**Figure 2. Differential G4 formation positively correlated with molecular phenotypes in Bloom Syndrome.** (**a**) Positive correlation between differential chromatin accessibility and differential gene expression in LCL-BS vs. LCL-WT. The red line indicates the fitted linear regression line. R is the Pearson correlation coefficient. ***, $P < 2\times10^{-16}$. (**b**) Positive correlation between differential chromatin accessibility and differential G4 formation in LCL-BS vs. LCL-WT. The red line indicates the fitted linear regression line. R is the Pearson correlation coefficient and ***, $P < 2\times10^{-16}$. (**c**) Distribution of differentially accessible chromatin regions and their fold changes stratified by the changes in G4 formation in LCL-BS vs. LCL-WT. Red dots denote the median of $log_2FC$. N.d., non-differential signals with p-adj ≥ 0.05. (**d**) Positive correlation between differential gene expression and differential G4 formation in LCL-BS vs. LCL-WT. The red line is the fitted linear regression line. *R* is the Pearson correlation coefficient. ***, $P < 2\times10^{-16}$. (**e**) Distribution of differentially expressed genes and their fold changes stratified by the changes in G4 formation in LCL-BS vs. LCL-WT. Red dots denote the median of $log_2FC$. (**f**) Enrichment of sister chromatid exchange (SCE) events at endogenous G4 forming sites in LCL-BS. Z-tests were carried out to compare observed overlap and expected overlap. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

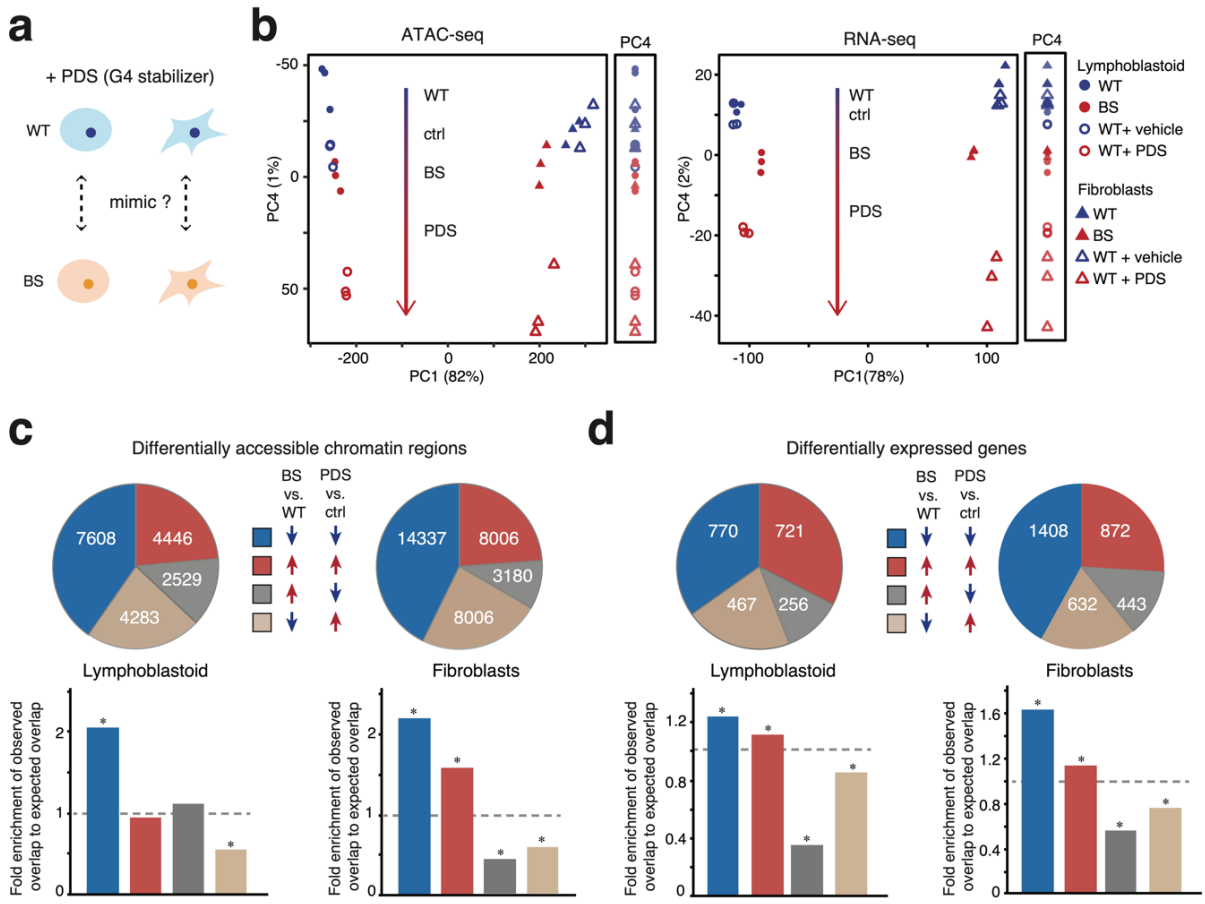



**Figure 3. G-quadruplex stabilization via PDS partially recapitulated chromatin accessibility and gene expression changes in Bloom Syndrome.** (**a**) Experimental design. (**b**) Principal component analysis of the genome-wide chromatin accessibility (left) and gene expression profiles (right) in samples across all cell types and conditions. (**c**) Overlap of differentially accessible chromatin regions (top) and their fold enrichment relative to expected (bottom) in Bloom Syndrome and under PDS treatment in LCL and Fib, respectively. *, hypergeometric tests, $P < 0.001$ (**Suppl. Table 2**). (**d**) Overlap of differentially expressed genes (top) and their fold enrichment relative to expected (bottom) in Bloom Syndrome and under PDS treatment in LCL and Fib, respectively. *, hypergeometric tests, $P < 0.001$ (**Suppl. Table 3**).

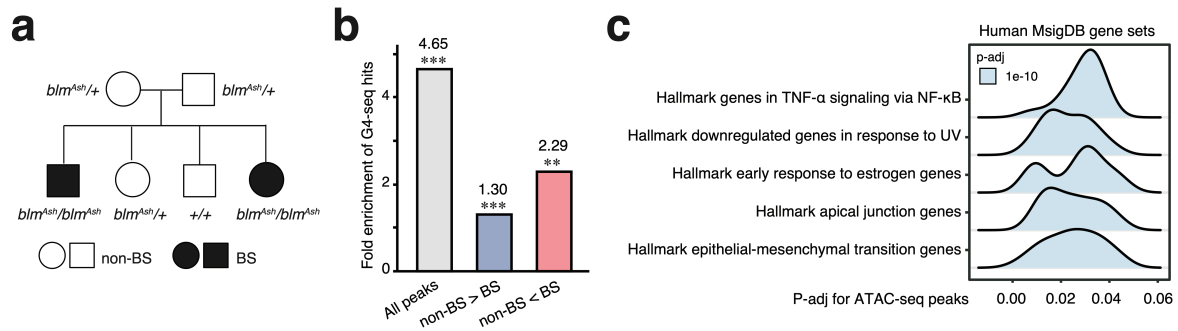



**Figure 4. More accessible chromatin regions in Bloom Syndrome individuals were more enriched for G4-seq hits.** (**a**) Pedigree of the Bloom Syndrome family. (**b**) Enrichment of G4-seq hits in differentially accessible chromatin regions. Z-test, **, $P < 0.01$; ***, $P < 0.001$. (**c**) Gene set enrichment analysis against hallmark pathways on the differentially accessible chromatin regions.

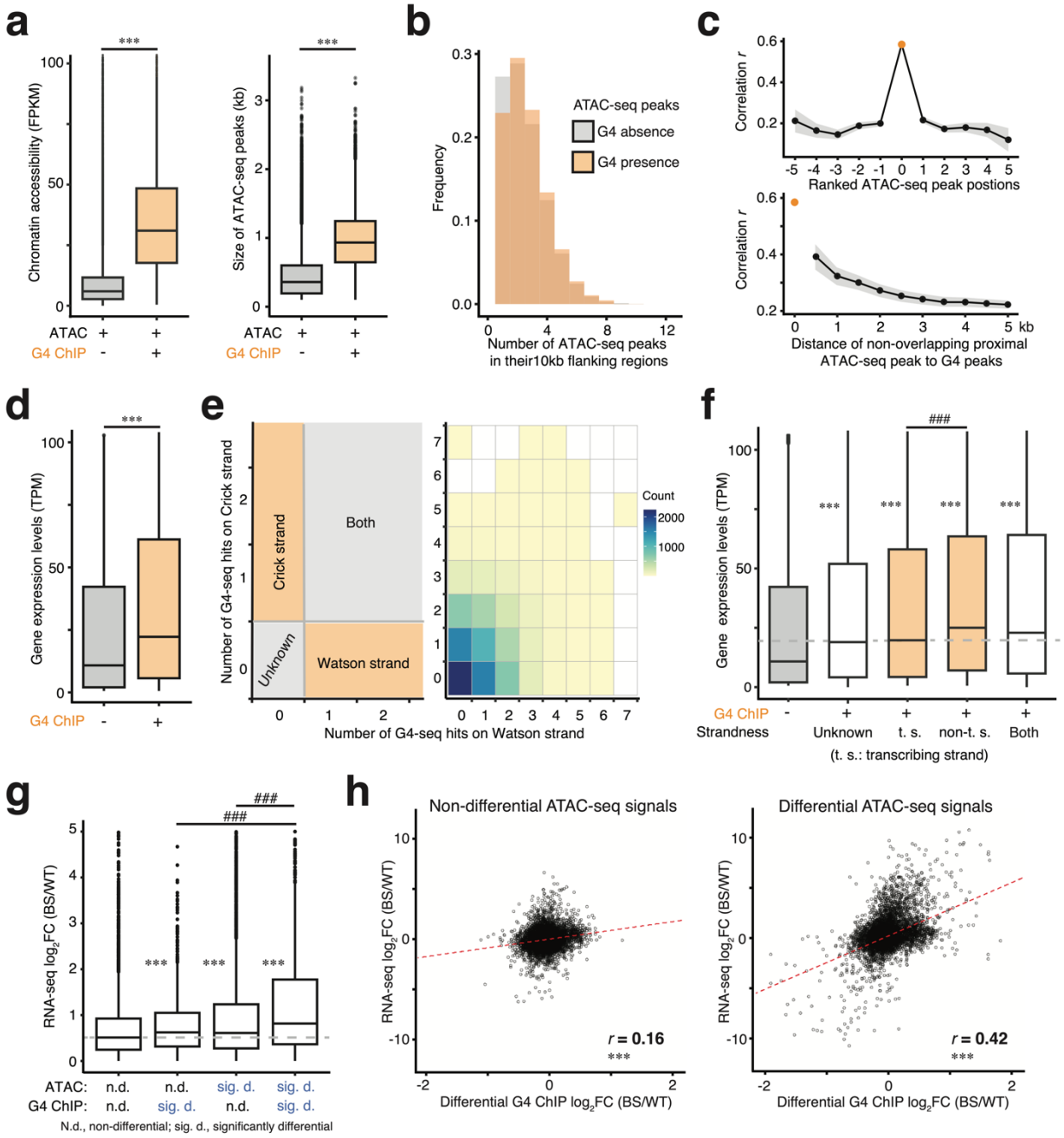



**Figure 5. G-quadruplexes plausibly regulated gene expression via regulating chromatin accessibility.** (**a**) ATAC-seq peaks overlapping G4 peaks are associated with higher chromatin accessibility and larger open chromatin sizes. (**b**) The presence of G4 peaks is associated with denser open chromatin regions. (**c**) Correlation between G4 ChIP-seq $log_2$FC and ATAC-seq $log_2$FC of G4-overlapping (orange dots) or non-overlapping proximal (black dots) ATAC-seq peaks. (**d**) Genes with G4 formation in the promoter displaying higher gene expression. (**e**) Inferring the strandedness of G4 peaks by intersecting with G4-seq hits. G4 in G4 peaks containing G4-seq hits only from Watson (or Crick) strand were inferred to form on Watson (or Crick) strand. (**f**) Genes exhibiting higher gene expression when G4 forms on the non-transcribing strand compared to G4 forming on the transcribing strand. Wilcoxon tests

were used. Groups with *P* values denoted by * were compared to the gene without G4 (the most left boxplot); ***, $P < 2\times10^{-16}$; ###, $P = 0.0008$. (**g**) Additive effects of differential chromatin accessibility and G4 formation on differential gene expression. Wilcoxon tests were used. Groups with P values denoted by * were compared to the gene without G4 (the most left boxplot); ***, $P < 1\times10^{-6}$; ###, $P < 2\times10^{-16}$. (**h**) The positive correlation between differential G4 formation and differential gene expression stratified by a significant change in ATAC-seq signature. The red line indicates the fitted linear regression line. R is the Pearson correlation coefficient and ***, $P < 2\times10^{-16}$.

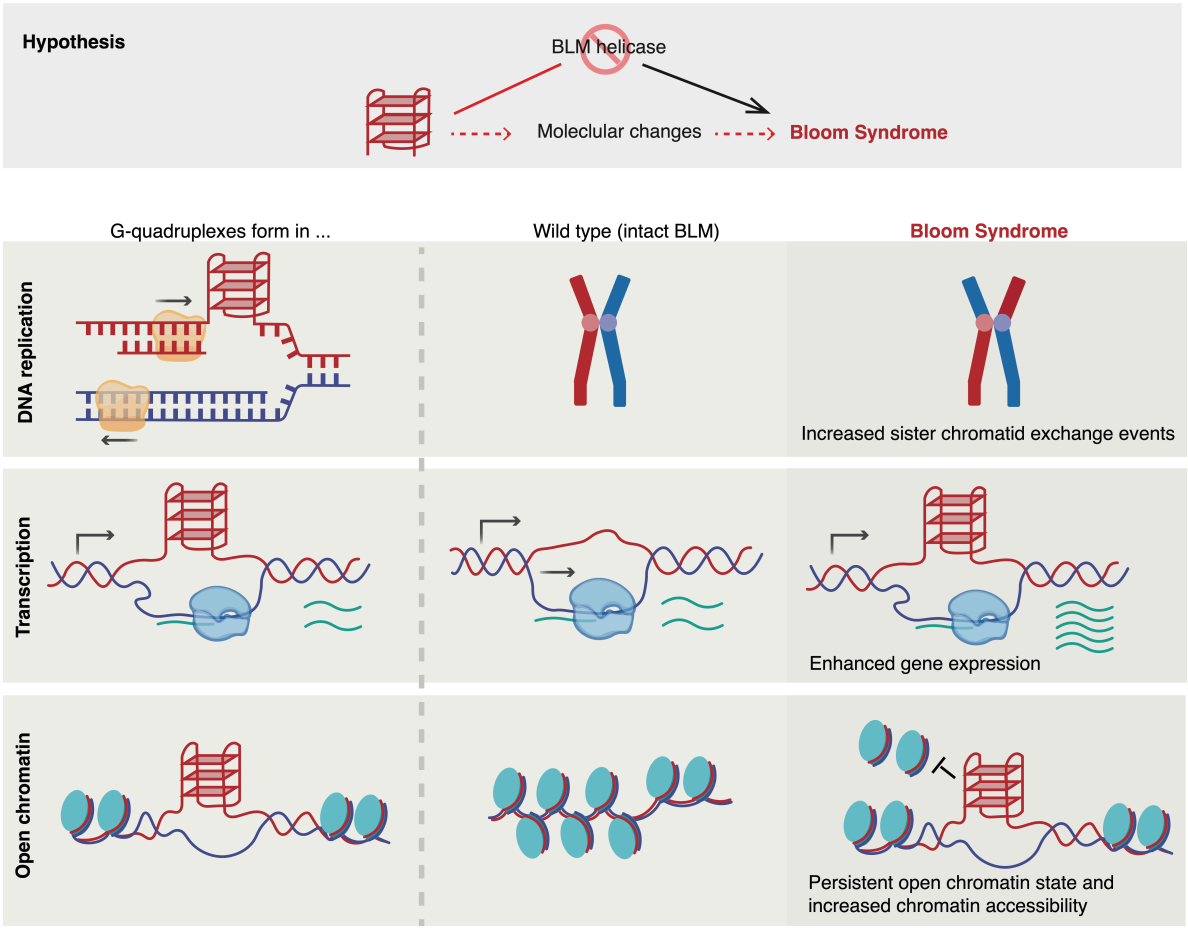
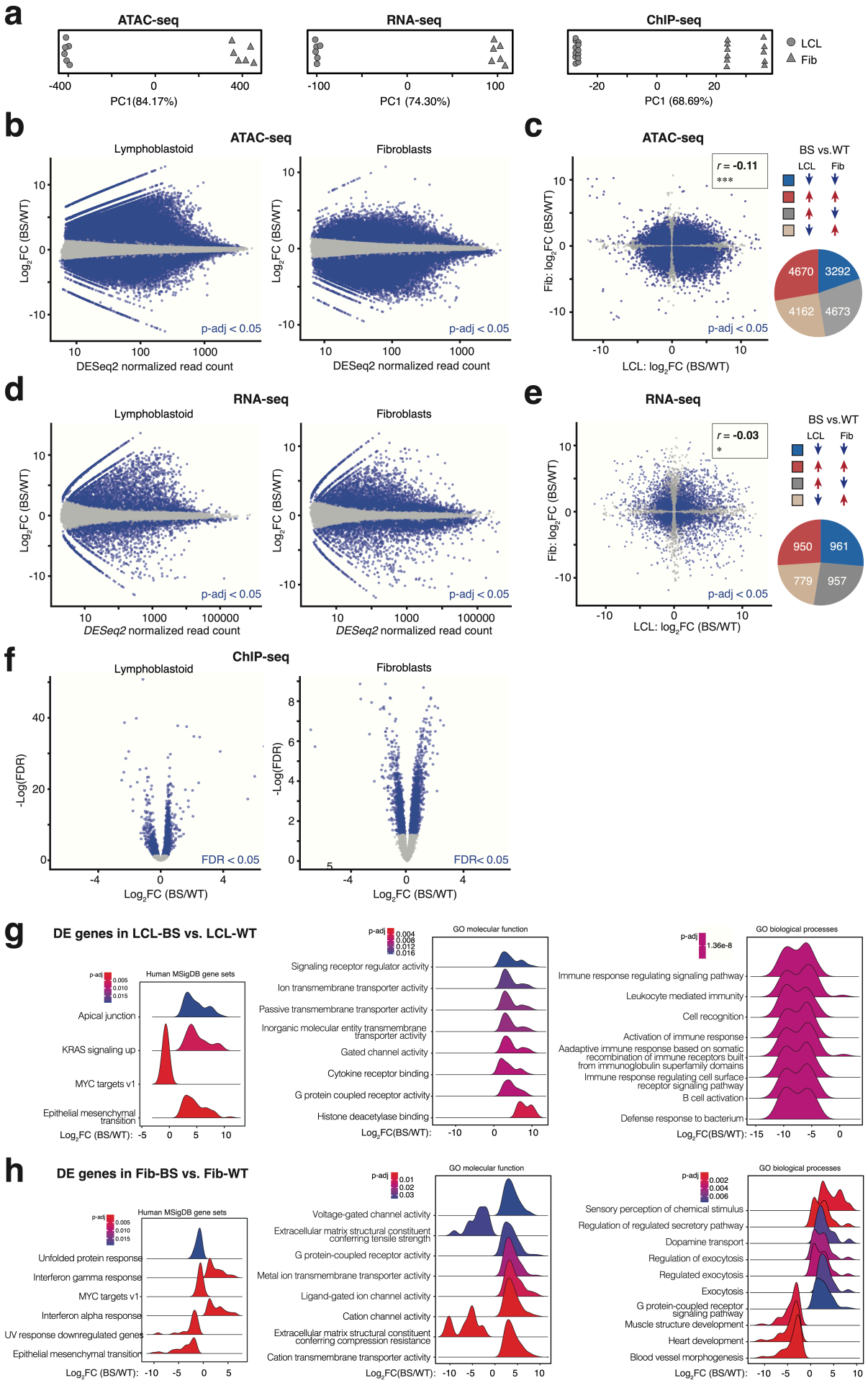
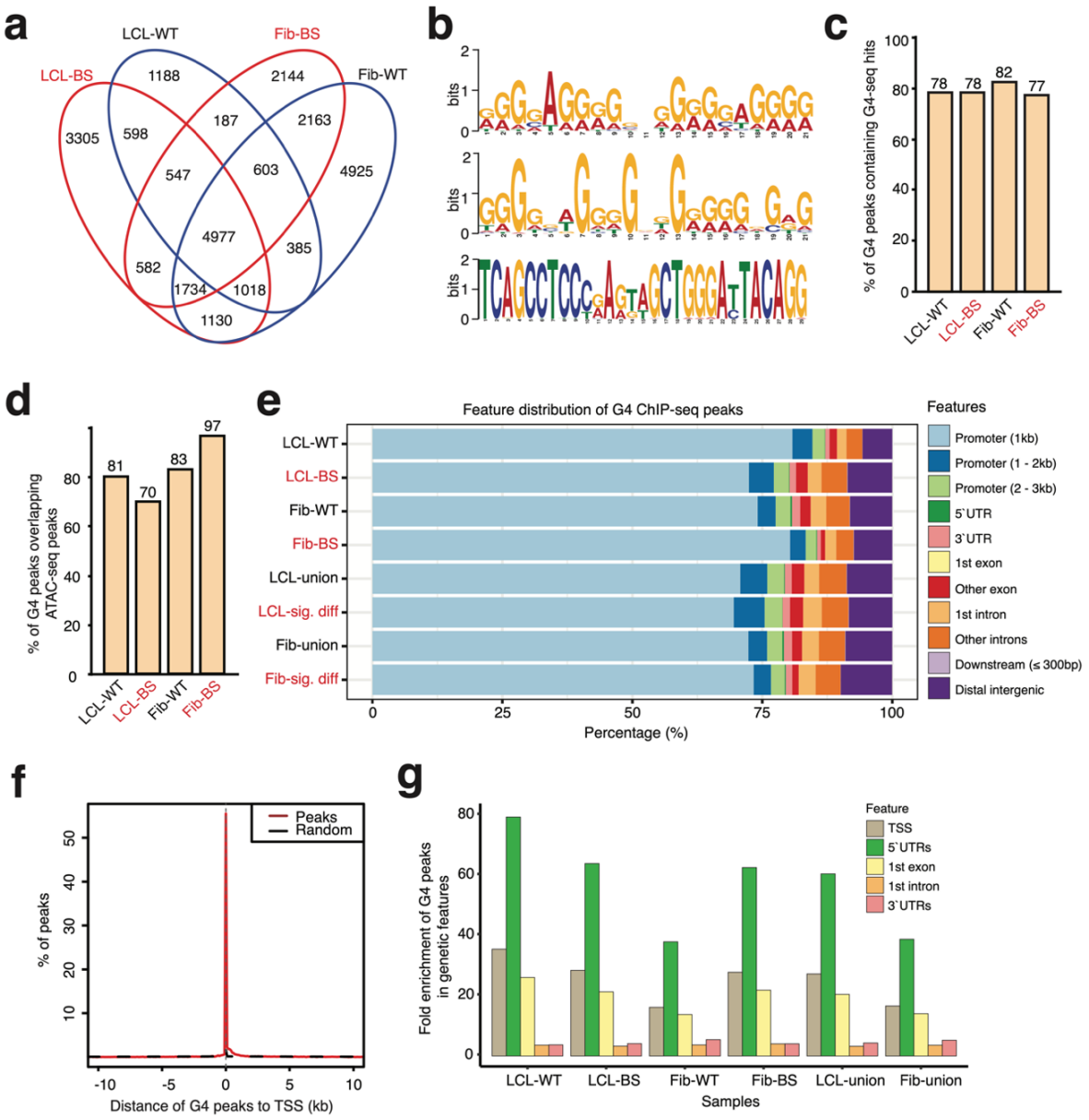


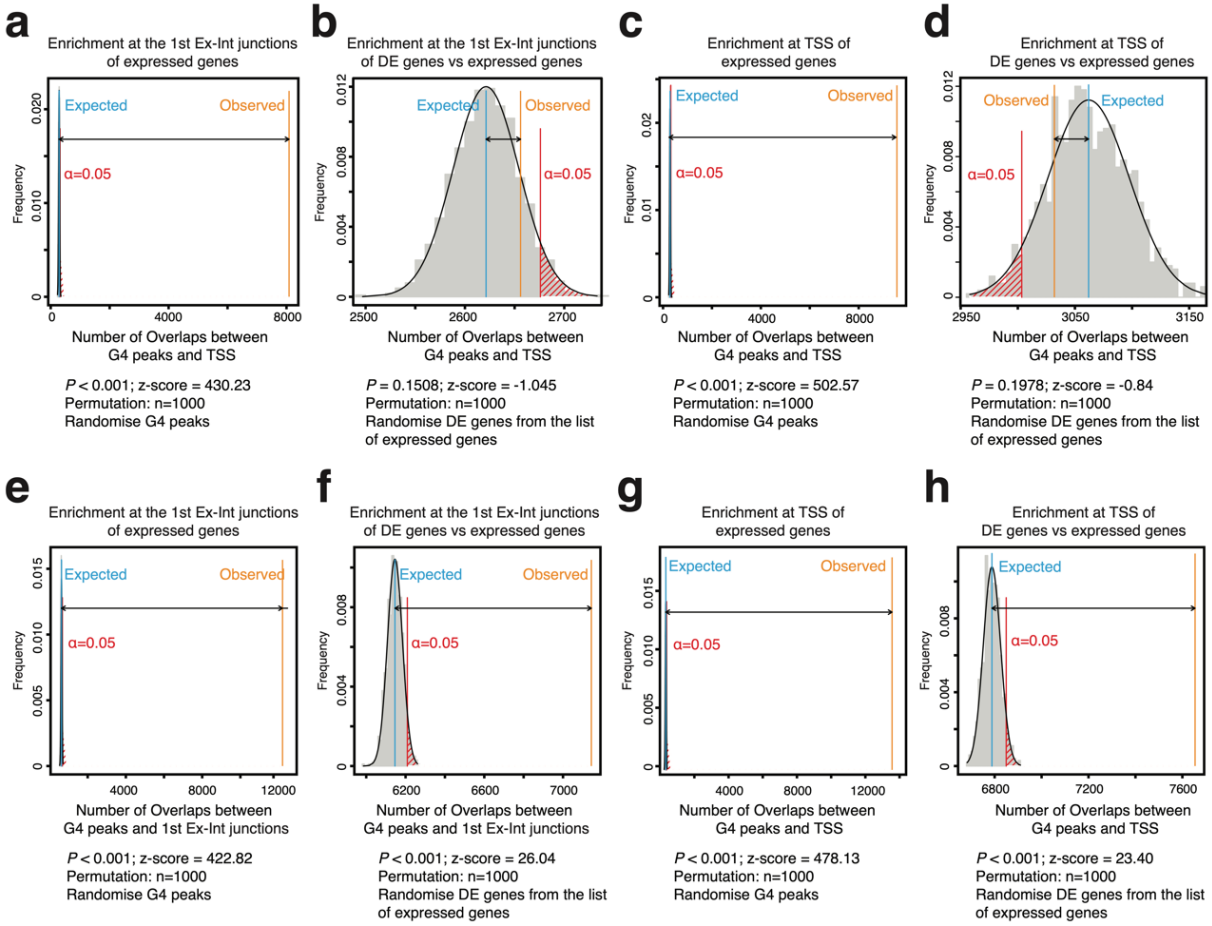

**Figure 6. Molecular model of G4 in the molecular etiology of Bloom Syndrome.**


a
ATAC-seq
RNA-seq
ChIP-seq
LCL
Fib
PC1(84.17%)
PC1 (74.30%)
PC1 (68.69%)
b
ATAC-seq
Lymphoblastoid
Fibroblasts
Log2FC (BS/WT)
p-adj < 0.05
DESeq2 normalized read count
c
ATAC-seq
r = -0.11
***
BS vs.WT
LCL
Fib
4670
3292
4162
4673
Fib: log2FC (BS/WT)
LCL: log2FC (BS/WT)
d
RNA-seq
DESeq2 normalized read count
e
RNA-seq
r = -0.03
*
950
961
779
957
f
ChIP-seq
-Log(FDR)
FDR < 0.05
FDR< 0.05
Log2FC (BS/WT)
g
DE genes in LCL-BS vs. LCL-WT
Human MSigDB gene sets
Apical junction
KRAS signaling up
MYC targets v1
Epithelial mesenchymal transition
GO molecular function
Signaling receptor regulator activity
Ion transmembrane transporter activity
Passive transmembrane transporter activity
Inorganic molecular entity transmembrane transporter activity
Gated channel activity
Cytokine receptor binding
G protein coupled receptor activity
Histone deacetylase binding
GO biological processes
Immune response regulating signaling pathway
Leukocyte mediated immunity
Cell recognition
Activation of immune response
Aadaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains
Immune response regulating cell surface receptor signaling pathway
B cell activation
Defense response to bacterium
h
DE genes in Fib-BS vs. Fib-WT
Unfolded protein response
Interferon gamma response
MYC targets v1
Interferon alpha response
UV response downregulated genes
Epithelial mesenchymal transition
Voltage-gated channel activity
Extracellular matrix structural constituent conferring tensile strength
G protein-coupled receptor activity
Metal ion transmembrane transporter activity
Ligand-gated ion channel activity
Cation channel activity
Extracellular matrix structural constituent conferring compression resistance
Cation transmembrane transporter activity
Sensory perception of chemical stimulus
Regulation of regulated secretory pathway
Dopamine transport
Regulation of exocytosis
Regulated exocytosis
Exocytosis
G protein-coupled receptor signaling pathway
Muscle structure development
Heart development
Blood vessel morphogenesis

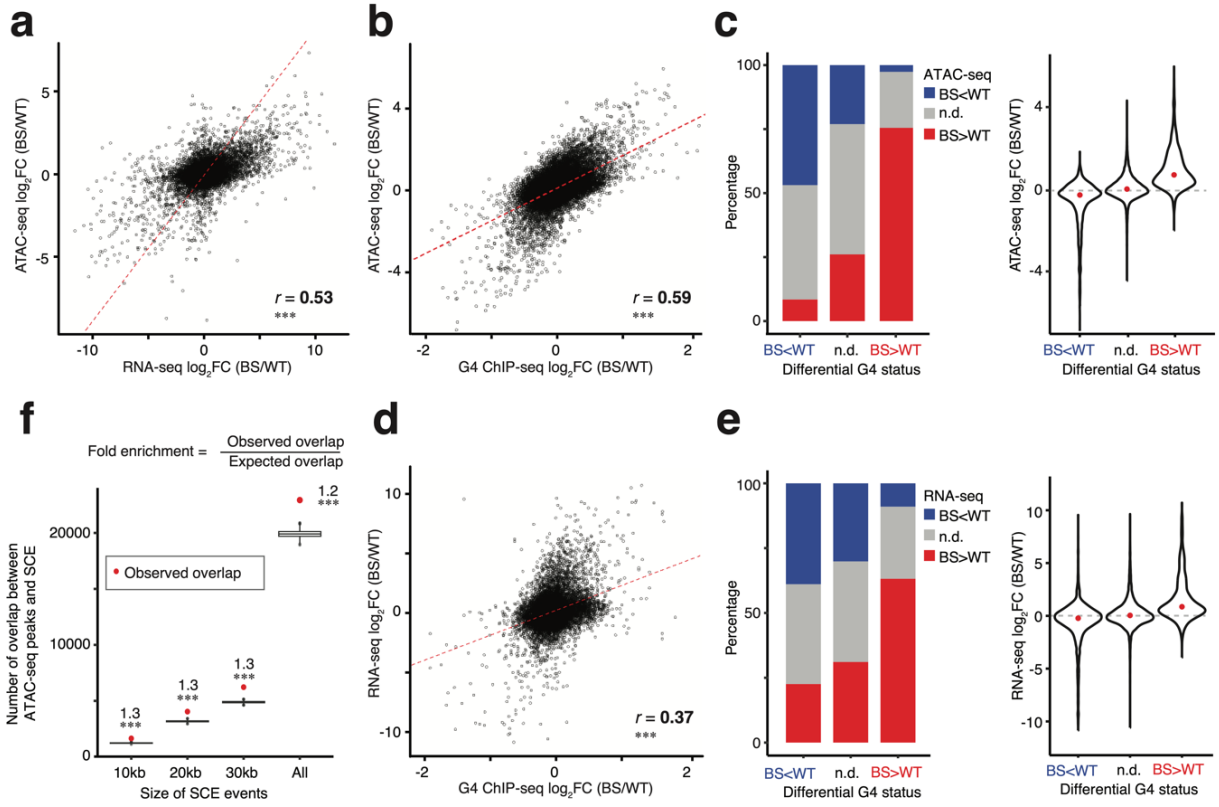
**Supplementary Figure 1. Changes in chromatin accessibility, gene expression and G4 formation in Bloom Syndrome cells.** (**a**) PC1 for ATAC-seq, RNA-seq and ChIP-seq. (**b**) MA-plot of differentially accessible ATAC-seq peaks in LCL-BS vs. LCL-WT (left) and Fib-BS vs. Fib-WT (right). Peaks with p-adj < 0.05 are denoted in blue and those with p-adj ≥ 0.05 in gray. (**c**) Comparison of differentially accessible ATAC-seq peaks in LCL-BS vs. LCL-WT and Fib-BS vs. Fib-WT. Peaks with significant chromatin accessibility changes in both comparisons are denoted in blue and the rest in gray (left). Blue and red arrows indicate significant decreases and increases in the specified contrast, respectively (right). (**d**) MA-plot of differentially expressed genes in the RNA-seq analysis. Left, LCL-BS vs. LCL-WT; right, Fib-BS vs. Fib-WT. Significantly differentially expressed genes with p-adj < 0.05 are denoted in blue and those with p-adj ≥ 0.05 in gray. (**e**) Comparison of differentially expressed genes in LCL-BS vs. LCL-WT and Fib-BS vs. Fib-WT. Significantly differentially expressed genes in both comparisons are denoted in blue and the rest in gray (left). Blue and red arrows indicate significant decreases and increases in the specified contrast, respectively (right). R is the Pearson correlation coefficient. *, $P < 0.05$; ***, $P < 2\times10^{-16}$. (**f**) Volcano plot of differentially G4-forming sites in ChIP-seq data analysis. Left, LCL-BS vs. LCL-WT; right, Fib-BS vs. Fib-WT. Peaks with FDR < 0.05 are denoted in blue and those with FDR ≥ 0.05 in gray. Gene ontology and Gene set enrichment analysis with the significantly differentially expressed genes ranked by fold change in LCL-BS vs. LCL-WT (**g**) and Fib-BS vs. Fib-WT (**h**).



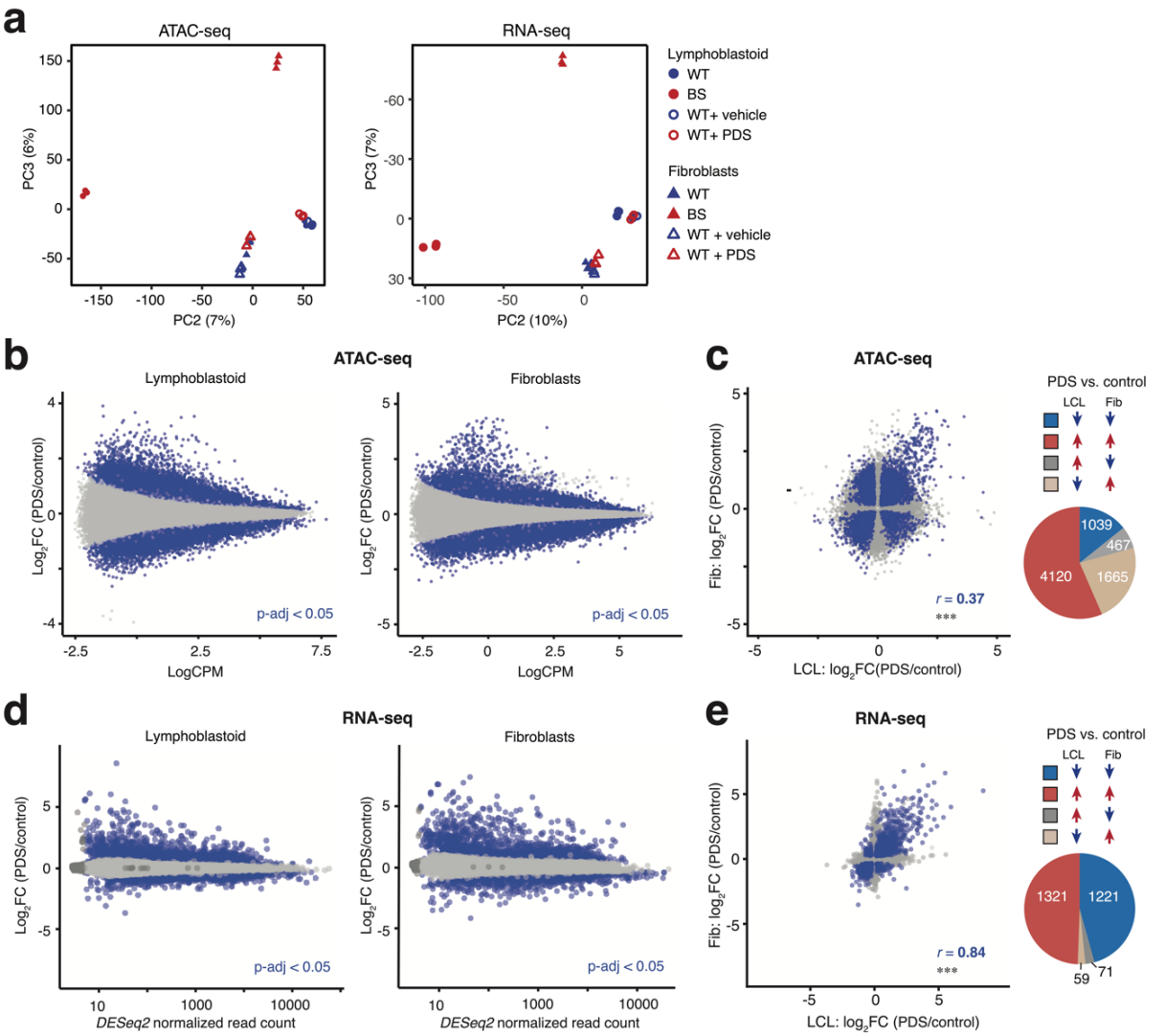

**Supplementary Figure 2. Features of endogenous G4 peaks.** (**a**) Representation (overview) of G4 peaks captured by G4 ChIP-seq in each analyzed sample. (**b**) Examples of enriched DNA motifs in G4 peaks. (**c**) Percentage of G4 peaks overlapping with G4-seq hits. (**d**) Percentage of G4 peaks overlapping open chromatin regions. (**e**) Distribution of G4 peaks in different genetic features. (**f**) G4 peaks are enriched at TSS. (**g**) Fold enrichment of G4 peaks in different genetic features. Fold enrichment is the ratio of the observed overlap between G4 peaks and the genetic features to the expected overlap estimated by permutation (see Methods).




**Supplementary Figure 3. Endogenous G4 peaks are not consistently enriched at the first exon-intron junctions of differentially expressed genes in Bloom Syndrome.** In LCL-BS vs. LCL-WT: enrichment of G4 peaks at the first exon-intron junctions (1st-Ex-Int) of expressed genes compared to random expectation (**a**), of differentially expressed genes compared to expressed gene (**b**); enrichment of G4 peaks at the TSS of expressed genes compared to random expectation (**c**), of differentially expressed (DE) genes compared to expressed genes (**d**).
In Fib-BS vs. Fib-WT: enrichment of G4 peaks at the 1st-Ex-Int of expressed genes compared to random expectation (**e**), of differentially expressed genes compared to expressed gene (**f**); enrichment of G4 peaks at the TSS of expressed genes compared to random expectation (**g**), of differentially expressed genes compared to expressed gene (**h**).

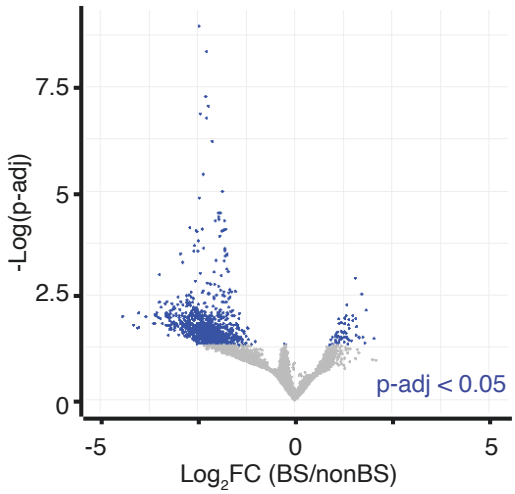



**Supplementary Figure 4. Differential G4 formation positively correlates with differential chromatin accessibility and differential gene expression in Bloom Syndrome.** For (**a**), (**b**) and (**d**), *r* is the Pearson correlation coefficient; ***, $P < 2\times10^{-16}$. (**a**) Positive correlation between differential chromatin accessibility and differential gene expression in Fib-BS vs. Fib-WT. The red line indicates the fitted linear regression line. (**b**) Positive correlation between differential chromatin accessibility and differential G4 formation in LCL-BS vs. LCL-WT. The red line indicates the fitted linear regression line. (**c**) Distribution of differentially accessible chromatin regions and their fold changes stratified by the changes in G4 formation in Fib-BS vs. Fib-WT. (**d**) Positive correlation between differential gene expression and differential G4 formation in Fib-BS vs. Fib-WT. The red line indicates the fitted linear regression line. (**e**) Distribution of differentially expressed genes and their fold changes stratified by the changes in G4 formation in Fib-BS vs. Fib-WT. (**f**) Enrichment of sister chromatid exchange (SCE) events in LCL-BS at open chromatin regions in LCL-BS. Z-tests were carried out to compare observed overlap and expected overlap. ***, $P < 0.001$.

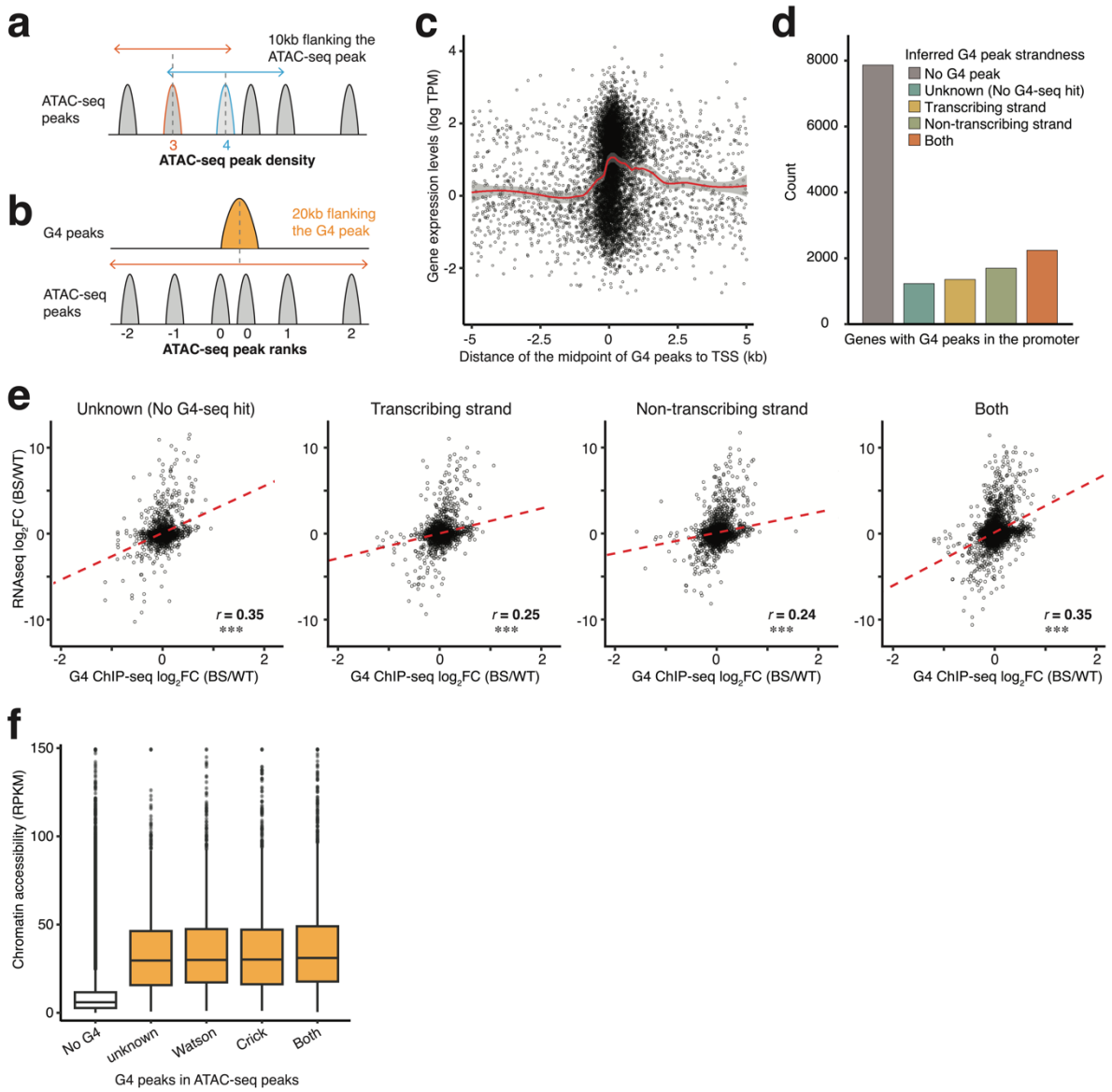



**Supplementary Figure 5. Changes in chromatin accessibility and gene expression caused by PDS treatment in wild-type cells.** (**a**) PCA on the global open chromatin landscape (left) and the gene expression profile (right) of WT (filled blue symbols), BS (filled red symbols), WT treated with vehicle (unfilled blue symbols), and WT samples treated with PDS (unfilled red symbols). (**b**) MA-plot of differentially accessible ATAC-seq peaks upon PDS treatment in LCL (left) and Fib (right). Significantly differentially expressed genes with p-adj < 0.05 are denoted in blue and those with p-adj ≥ 0.05 in gray. (**c**) Comparison of differentially accessible ATAC-seq peaks upon PDS treatment in LCL and Fib. Significantly differential signals in both cells are denoted in blue and the rest in gray (left). Blue and red arrows indicate significant decreases and increases in the specified contrast, respectively (right). R is the Pearson correlation coefficient for the blue data points. ***, $P < 2\times10^{-16}$. (**d**) MA-plot of differentially expressed genes upon PDS treatment in LCL (left) and Fib (right). Significantly differentially expressed genes with p-adj < 0.05 are denoted in blue and those with p-adj ≥ 0.05 in gray. (**e**) Comparison of differentially expressed genes upon PDS treatment in LCL and Fib. Significantly differential signals in both cells are denoted in blue and the rest in gray (left). Blue and red arrows indicate significant

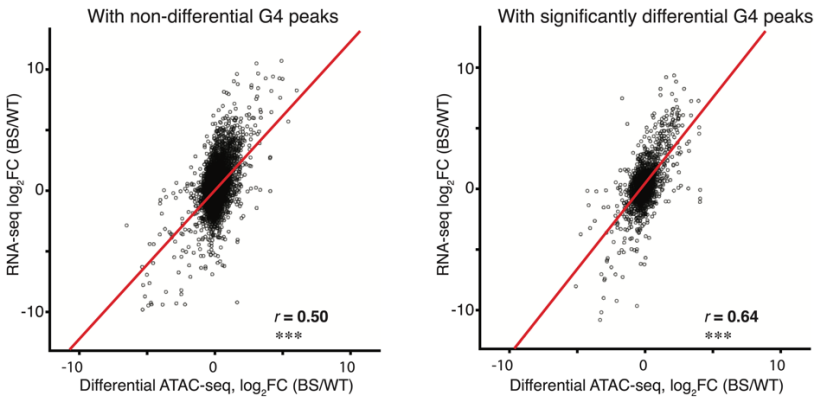
decreases and increases in the specified contrast, respectively (right). R is the Pearson correlation coefficient for the blue data points. ***, $P < 2\times10^{-16}$.



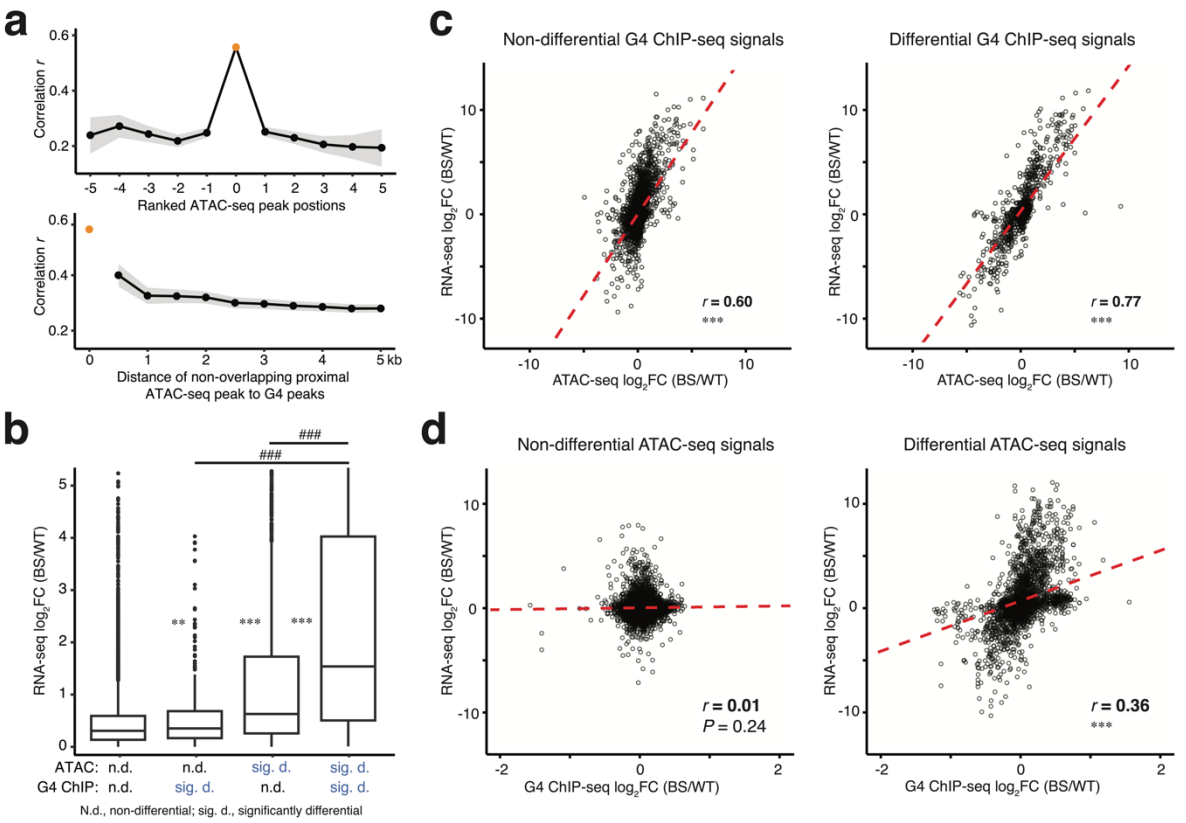

**Supplementary Figure 6.** Differentially accessible chromatin regions in Bloom Syndrome compared to the non-Bloom Syndrome individuals. Significantly differentially expressed genes with p-adj < 0.05 are denoted in blue and those with p-adj ≥ 0.05 in gray.




**Supplementary Figure 7. Impacts of differential G4 formation on differential chromatin accessibility and differential gene expression.** (**a**) Schematic representation of estimating ATAC-seq peak density by counting the number of ATAC-seq peaks in 10 kb flanking the midpoint of an ATAC-seq peak. (**b**) Schematic representation of ranked ATAC-seq peak positions proximal to G4-peaks. Overlapping ATAC-seq peaks are ranked 0. Upstream and downstream ATAC-seq peaks have negative and positive ranks, respectively. (**c**) The gene expression levels tending to decrease as the G4 peaks form further away from the TSS. The red line denotes the LOESS fitted curve for the data and the grey area depicts the 95% confidence interval for the fitted curve. (**d**) Number of genes with inferred strandedness of G4. (**e**) Correlation between G4 formation and gene expression changes stratified by the strandedness of the G4 relative to the transcribing strand. The red line indicates the fitted linear regression line. R is the Pearson correlation coefficient. ***, $P < 2\times10^{-16}$. (**f**) Chromatin accessibility in ATAC-seq peaks stratified by inferred strandedness of the G4 peaks.




**Supplementary Figure 8.** Correlation between differential chromatin accessibility and differential gene expression stratified by the differential status of G4 peaks. The red line indicates the fitted linear regression line. R is the Pearson correlation coefficient. ***, $P < 2\times10^{-16}$.




**Supplementary Figure 9. Molecular insights into G4 possibly regulating gene expression via modulating chromatin accessibility in lymphoblastoid.** (**a**) Correlation between G4 ChIP-seq $log_2$FC and ATAC-seq $log_2$FC of G4-overlapping or proximal ATAC-seq peaks. (**b**) Synergistic effects of differential chromatin accessibility and G4 formation on differential gene expression in LCL-BS vs. LCL-WT. Wilcoxon tests were used. Groups with *P* values denoted by * were compared to the gene without G4 (the most left boxplot); **, $P = 0.003$, $P < 2\times10^{-16}$; ***, $P < 2\times10^{-16}$; ###, Wilcoxon test, $P < 2\times10^{-16}$. (**c**) Correlation between differential gene expression and differential chromatin accessibility for genes with (right) and without (left) significantly unaltered G4 formations. The red line indicates the fitted linear regression line. *R* is the Pearson correlation coefficient and *** denotes $P < 2\times10^{-16}$. (**d**) Correlation between differential gene expression and differential G4 formations for genes with (right) and without (left) significantly altered chromatin accessibility. R is the Pearson correlation coefficient and *** denotes $P < 2\times10^{-16}$. The red line indicates the fitted linear regression line.

**Supplementary Table 1. Enrichment of sister chromatin exchange events in G4 peaks and ATAC-seq peaks.**

| | Number of overlaps with G4 peaks | | | | Number of overlaps with ATAC-seq peaks | | | |
|---|---|---|---|---|---|---|---|---|
| | All SCEs | SCE (<10kb) | SCE (<20kb) | SCE (<30kb) | All SCEs | SCE (<10kb) | SCE (<20kb) | SCE (<30kb) |
| Observed | 2927 | 209 | 530 | 785 | 22934 | 1648 | 4042 | 6225 |
| Expected | 2644.5 ±109.1 | 176.5 ±14.8 | 438.5 ± 25.3 | 665.7 ± 33.1 | 19906.8 ± 320.0 | 1231.6 ± 38.2 | 3177.4 ± 70.5 | 4891.5 ± 88.5 |
| Fold enrichment | 1.11 | 1.18 | 1.21 | 1.18 | 1.15 | 1.34 | 1.27 | 1.27 |
| Z score | 2.55 | 2.20 | 3.62 | 3.61 | 9.73 | 10.91 | 12.27 | 15.06 |
| *P* | 1.40 E-02 | 2.20 E-02 | 2.00 E-03 | 9.99 E-04 | 9.99 E-04 | 9.99 E-04 | 9.99 E-04 | 9.99 E-04 |

Expected number of overlaps is estimated by permutation (n=1000). The number is represented as mean ± standard deviation.

**Supplementary Table 2. Hypergeometric tests of enrichment and depletion of overlapping differentially accessible ATAC-seq peaks upon BLM deficiency and PDS treatment.**

| DA peak categories | Observed | Expected by chance | Fold enrichment | Log *P* (Hypergeometric test) |
|---|---|---|---|---|
| LCL: BS- PDS- | 7608 | 3707 | **2.05** | -2537.96 |
| LCL: BS+ PDS+ | 4446 | 4715 | 0.94 | -5.24156E-07 |
| LCL: BS- PDS+ | 4283 | 3858 | 1.11 | -3.11043E-16 |
| LCL: BS+ PDS- | 2529 | 4530 | **0.56** | -756.1775 |
| Fib: BS- PDS- | 14657 | 6726 | **2.18** | -7176.127 |
| Fib: BS+ PDS+ | 8278 | 5799 | **1.43** | -1154.135 |
| Fib: BS- PDS+ | 3491 | 5304 | **0.66** | -858.8632 |
| Fib: BS+ PDS- | 2418 | 6152 | **0.39** | -1765.167 |

BS+ and BS -: significantly increased or decreased signal in BS compared to WT; PDS+ and PDS -: significantly increased or decreased signal in PDS compared to control. Log (0.05) = -1.30; Log(*P*) > -1.30 for $P < 0.05$. Fold enrichment values in bold denotes $P < 0.05$ of the corresponding hypergeometric tests.

**Supplementary Table 3. Hypergeometric tests of enrichment and depletion of overlapping differentially expressed genes upon BLM deficiency and PDS treatment.**

BS+ and BS -: significantly increased or decreased signal in BS compared to WT; PDS+ and

| DE gene catogories | Observed | Expected by chance | Fold enrichment | *P* (Hypergeometric test) |
|---|---|---|---|---|
| LCL: BS- PDS- | 770 | 622 | **1.24** | 2.15E-13 |
| LCL: BS+ PDS+ | 721 | 646 | **1.12** | 1.23E-131 |
| LCL: BS- PDS+ | 467 | 548 | **0.85** | 3.25E-05 |
| LCL: BS+ PDS- | 256 | 733 | **0.35** | 0.000135306 |
| Fib: BS- PDS- | 1408 | 859 | **1.64** | 3.89E-133 |
| Fib: BS+ PDS+ | 872 | 760 | **1.15** | 6.69E-66 |
| Fib: BS- PDS+ | 632 | 824 | **0.77** | 4.02E-20 |
| Fib: BS+ PDS- | 443 | 793 | **0.56** | 7.78E-08 |

PDS -: significantly increased or decreased signal in PDS compared to control. Log (0.05) = -1.30; Log(*P*) > -1.30 for $P < 0.05$. Fold enrichment values in bold denotes $P < 0.05$ of the corresponding hypergeometric tests.

**Supplementary Table 4. Mutations in *BLM* gene and medical records of the Bloom Syndrome family.**

| Sample | Sex | Relation to proband | Syndrome status | *BLM alleles* | Age at donation (month) | Neoplasm |
|---|---|---|---|---|---|---|
| FBL131 | M | Father | Carrier | *blm*$^{Ash}$/+ | 456 | |
| FBL132 | F | Mother | Carrier | *blm*$^{Ash}$/+ | 446 | |
| FBL305 | M | Patient 1 | Affected | *blm*$^{Ash}$/ *blm*$^{Ash}$ | 19 | Wilm's tumor at 36 months, AML at 72 months, myelodyplastic syndrome at 72 months |
| FBL306 | F | Patient 2 | Affected | *blm*$^{Ash}$/ *blm*$^{Ash}$ | 63 | None |
| FBL301 | M | Brother | Wild-type | +/+ | 171 | |
| FBL302 | F | Sister | Carrier | *blm*$^{Ash}$/+ | 151 | |

*blm*$^{Ash}$*: c.2207_2212delATCTGAinsTAGATTC in exon10 in BLM.*

**Supplementary Table 5. Primer sequences used in G4 ChIP qPCR.**

| Primer | Sequence | G4-ChIP peak | G4 motifs |
|---|---|---|---|
| KIF14_G4_for | CGGTAGCCGTCTCTGAATG | + | + |
| KIF14_G4_rev | CTTTAGCAGAACCCGAGGAG | | |
| RPA3_G4_for | CGGAAGTTGACAGATACAGGG | + | + |
| RPA3_G4_rev | GATCGCAGAAAGGTAGTCTCAG | | |
| GAPDH_G4_for | GCTACTAGCGGTTTTACGGGCG | + | + |
| GAPDH_G4_rev | TGCGGCTGACTGTCGAACAGG | | |
| HTR6_G4_for | GGCGATTTGTCCAATATTTCCC | - | + |
| HTR6_G4_rev | CTGTGACCTGCCCTTATCC | | |
| ESR1_for | GAAACAGCCCCAAATCTCAA | - | - |
| ESR1_rev | TTGTAGCCAGCAAGCAAATG | | |
| TMCC1_for | GTGGTACACTGCCTACAGTATT | - | - |
| TMCC1_rev | GTATAACGCCTGGGCTATGT | | |

# Methods

### Cell culture

Sex-matched and roughly age-matched fibroblast cell lines from Bloom syndrome and healthy donors were obtained from Coriell Institute (lymphoblastoid cell lines, LCL: GM16375, AG14980; fibroblast cell lines: GM08505, GM00637). LCLs were cultured in RPMI 1640 medium (Gibco, Cat. #21875) supplemented with 15% fetal bovine serum (FBS, Gibco, UK, Cat. #10500), 1% Penicillin-Streptomycin (Thermo Scientific, Gibco, Cat. #15140122) and 1% GlutaMAX-I (Gibco, UK, Cat. #35050). Fibroblast lines were cultured in 1× DMEM (Thermo Scientific, Gibco, Cat. #11960) supplemented with 1% minimum essential medium non-essential amino acids (MEM NEAA; Thermo Scientific, Gibco, Cat. #11140) and 1% P/S and 10% or 15% fetal bovine serum (FBS; Thermo Scientific, Gibco, Cat. #10500064), respectively.

To collect cells for different experiments, fibroblasts were first trypsinised (Trypsin-EDTA solution, Sigma-Aldrich, Cat. #T4049), pelleted at 300 g for 5 min, and washed once with cold 1x Dulbecco's phosphate-buffered saline (PBS, Thermo Scientific, Gibco, Cat. #14190). Lymphoblastoid cells were pelleted at 300 g for 5 min and washed once with cold 1x PBS.

### Small molecule treatment

Pyridostatin (PDS, Sigma-Aldrich, Cat. #SML2690) stocks (2.5 mM) were prepared by dissolving it in DMSO (Sigma-Aldrich, Cat. #D2650). For LCLs, wildtype cells from cell line AG14980 were treated with 10 µM PDS vs. 0.4% DMSO (control) for 24 h at the concentration of 1 million cells per ml. Cells were then cultured with gentle rotation at 100 rpm to ensure equal exposure of cells to the treatment. For fibroblasts, wild-type cells (GM00637) at 60% confluency were treated with 10 µM PDS vs. 0.4% DMSO (referred as control) for 24 h.

### ATAC-seq library preparation

***Cell lines.*** Cells were collected as described above and counted (automated cell counter Countesss 3, Thermo Scientific; Countess™ Cell Counting Chamber Slides Cat. # C10283). For each reaction 50, 000 cells were subjected to ATAC-seq preparation essentially as described in Corces, M. *et al.* [75] with modified tagmentation reaction. Nuclei were resuspended in 50 µl of tagmentation mix consisting of 10 µl 5× TAPS-DMF buffer (50 mM TAPS, 25 mM $MgCl_2$, 50% (v/v) DMF), 3 µl Tn5 transposase [76], 16.5 µl 1x PBS, 0.25 µl 2% Digitonin, 0.5 µl 10% (v/v) Tween 20 and 19.75 µl $H_2O$. DNA purified from the transposed product was then amplified via PCR (10 cycles) using Q5 High-Fidelity DNA Polymerase (New England Biolabs, Cat. #M0491L) and indexed primers N5xx and N7xx (synthesized by Integrated DNA Technologies) (Supplementary Table).

***BS family samples.*** The procedure is the same as described above for lymphoblastoid exception for a longer cell lysis incubation on ice (10 min) and pelleting nuclei at 1000 g for 10 min.

### RNA extraction and RNA-seq library preparation

***RNA extraction.*** Cell collection was performed as described above. The collected cells were lysed in 1 ml of Trizol (Sigma, Cat. #15596026) at room temperature (RT) for 5 min. Subsequently, 200 µl of chloroform (Thermo, Cat. #J67241.AP) was added, and the samples

were centrifuged at 12,000 g for 15 min at 4°C to separate the phases. Prior to centrifugation, the samples were vortexed for 15 seconds and incubated at RT for 3 min. The upper phase, containing RNA, was carefully transferred to a new tube and mixed with 500 µl of isopropanol. The RNA was then precipitated at -20°C for 2 hours and subsequently pelleted at 12,000 g for 30 min at 4°C. The supernatant was discarded, and the RNA pellet was washed with 1 ml of ice-cold 75% ethanol. After centrifugation at 9,000 g for 5 min at 4°C, the ethanol wash step was repeated, and residual ethanol was aspirated using a pipette. The RNA pellet was air-dried by leaving the tubes open on the counter for approximately 5 min.The air-dried RNA pellet was resuspended in 50 µl of DNase reaction mix (5 µl of 10X Reaction Buffer for Turbo DNase, 1 µl of TURBO™ DNase (Invitrogen, Cat. #AM2238), 0.5 µl of recombinant RNase inhibitors (Jena Bioscience, Cat. #PCR-392S), and 43.5 µl of RNase-free water). The samples were incubated at 37°C for 25 min with gentle agitation at 550 rpm to carry out DNase treatment. For reextraction of RNA, 150 µl of $H_2O$ and 200 µl of Phenol/Chloroform/Isoamylalcohol were added. The samples were vortexed for 15 s and then centrifuged at 13,000 rpm for 2 min at 4°C. The upper phase was transferred to a new tube and mixed with 20 µl of 3M Na-acetate pH 5.5 and 200 µl of isopropanol. The samples were incubated at -20°C for 1 hour to allow for the second RNA precipitation. The RNA was then pelleted, and the washing procedure was repeated as described in the initial precipitation step. The RNA pellet obtained after the final precipitation steps was dissolved in H2O.

***RNA Quality assessment and library preparation.*** The quality of the RNA samples was assessed using Bioanalyzer RNA 6000 Nano assay (Agilent, Cat. # 5067-1511). Only RNA samples with a RIN (RNA Integrity Number) score exceeding 7 were selected for subsequent steps. RNA-seq library preparation was performed on 1 µg RNA using a poly-A mRNA enrichment kit (New England Biolabs, Cat. #E7490) and RNA-seq library preparation kit (New England Biolabs, Cat. #E7760S), following manufacturer's recommendations.

**G-quadruplex ChIP-seq protocol and library preparation**

**Chromatin preparation for fibroblasts and LCL.** Chromatin preparation was performed essentially as previously described with a few modifications [27]. About 30 Mil ~ 35 Mil LCL cells were collected and fixed with 0.5% formaldehyde (Thermo Scientific, Cat. #28906) in cell culture medium at RT for 6 min with gentle shaking. Cells were resuspended and treated in 1.5ml hypotonic buffer (Chromatrap, Cat. #100008) for 10 min. Nuclei were then lysed in 600ul nuclei lysis buffer (Chromatrap, Cat. #100008) and sheared in a 2ml sonication tube (Tube & Cap 12x24mm, Cat. # 520056) on a Covaries machine (Covaris S220) with the following parameters: *water level=15, duty cycle=18, intensity =7, cycle per burst = 200*. Each chromatin sample underwent shearing for a total of 15 min, with a 1 min pause for every 5 min of shearing.

For fibroblasts, cells from two full 15cm dishes were initially rinsed once in 1× PBS and subsequently fixed in the dish with 1% formaldehyde in cell culture medium at room temperature for 8.5 min. Nuclei were then lysed in 25 µl lysis buffer and sheared in small tubes (microTUBE AFA Fiber Pre-Slit Snap-cap 6×16mm, Covaris, Cat. #520045) using an S220 instrument (Covaris). The shearing parameters were as follows: *water level = 15*, *duty cycle = 15*, *intensity = 6*, and *cycle per burst = 200*. Each chromatin sample underwent shearing for a total of 3 min, with a 30-second pause for every one min of shearing.

***G-quadruplex chromatin immunoprecipitation.*** For each sample, 2 biological replicates were prepared, each in turn divided into 3 technical replicates of immune-precipitation, except

for the wild-type fibroblasts, for which the second biological replicate was performed with only 2 technical replicates due to the limited amount of BG4 antibody. The IP reaction was performed essentially as described previously [27], with the following modifications. During IP, chromatin was incubated with the BG4 antibody (Absolute antibody, Cat. #Ab00174-30.146) for an extended duration of 2 hours, and double the amount of anti-FLAG M2 magnetic beads (Sigma-Aldrich, Cat. #M8823; equivalent to 10 ul of original beads) were used in the pull-down. To purify DNA from both the IP samples (DNA pulled down by the beads) and the input samples (sheared chromatin without going through IP steps) after 5 times of washing the beads with WASH buffer (100 mM KCl, 0.1% (v/v) Tween 20, 10 mM Tris, pH 7.4) to remove residual nonspecific bound chromatin, 75 µl of reverse-crosslinking buffer (0.2% SDS, 1× TE, and 50 mM NaCl) was introduced. The samples were incubated at 37°C for one hour, followed by an overnight incubation at 65°C. After an additional hour of proteinase K digestion at 65°C, DNA purification was carried out using a MinElute kit (QIAGEN, Cat. #28006). The subsequent steps for library preparation were carried out with the DNA Thruplex kit (Takara, Cat. #R400674, Cat. #R400665) following the manufacturer's protocols.

***G-quadruplex ChIP-qPCR and library preparation***. Prior to library preparation, we used 2X CFX SYBR Mix (Applied Biosystems, Cat. #4472942) and a set of primers targeting known G4 positive and negative control sites to perform qPCR to evaluate G-quadruplex enrichment from IP vs. Input. Only samples with at least 5-fold enrichment of G4 positive sites (KLF14, GAPDH) to G4 negative sites (TMCC1) (**Supple. Table 5**) [27].

**Sequencing and data analysis**

One biological replicate of G4 ChIP-seq libraries from GM08505 and GM00637 were sequenced on a Hiseq3000 platform (paired-end sequencing of 150 nt each; Illumina Inc.; service provider: Genome center in the Max Planck Institute for Biology, Tübingen). The rest of the samples were sequenced on a NovaSeq6000 platform (paired-end sequencing of 150 nt each; Illumina Inc., San Diego, etc.; service provider: GENEWIZ, Leipzig)

**ATAC-seq data analysis**

***Read alignment and read filtering.*** Raw FASTQ reads were extracted with *bcl2fastq* (version 2.20). Tn5 and TruSeq adapter sequences, G repeats due to two-color base calling errors and reads shorter than 20 bp were trimmed and removed using *cutadapt* (version 4.0) [77]. Reads were then aligned to the human reference genome version hg38 using *bwa mem* (version: 0.7.17-r1188). Duplicates were marked and removed using *Picard* (https://broadinstitute.github.io/picard/, version 2.18.25). Prior to calling peaks, reads mapped to mitochondrial, reads with a mapping quality lower than 20 and reads in hg38 blacklisted regions (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38human/hg38.blacklist.bed.gz) were discarded.

***Mapping and peak calling for ATAC-seq.*** Raw FASTQ reads were extracted with bcl2fastq (version 2.20). Tn5 and Truseq adapter sequences, G repeats due to sequencing artefacts, and reads shorter than 20 bp were trimmed and removed using cutadapt (version 4.0) [77]. Reads were then aligned to human reference genome version hg38 or composite genome of hg38 and EBV using bwa mem (Version: 0.7.17-r1188) [78]. Duplicates were marked and removed using Picard (version 2.18.25) ("Picard Toolkit." 2019. Broad Institute, GitHub

Repository. https://broadinstitute.github.io/picard/). Prior to calling peaks, mitochondrial reads, reads with a mapping quality lower than 20 and reads in hg38 blacklisted regions (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38human/hg38.blacklist.bed.gz) were discarded.

ATAC-seq peak calling was performed using MACS2 (version 2.1.1.20160309) with *callpeak –format BAMPE --nomodel –min-length 100 narrowPeak* parameters [53]. For a given cell type, two types of ATAC-seq peak sets were generated. Using LCL as an example, to identify differentially accessible regions in each contrast (BS vs. WT or PDS vs. control), a local peak set was called by pooling an equal number reads from each sample. To identify peaks that are differentially accessible in BS and PDS-treated conditions, equal number of reads from WT, BS, control and PDS-treated samples from the same cell type was pooled to call a global peak set, which was further supplemented by unique peaks in the local peak sets from BS vs. WT or PDS vs. DMSO contrasts.

***Differential analysis.*** To identify differential signals in BS vs WT, number of fragments in each ATAC-seq peak was counted with htseq-count (HTSeq version 0.9.1) [79]. Subsequently, the count matrix was normalized and differential analysis was carried out ini*DESeq2* (version 1.30.1) to identify significantly differential signals (adjusted p-value < 0.05) [47,80]. In the case of applying copy number normalization, the count matrix was adjusted before data normalization as described in the following section. However, we noticed that the normalization method in *DESeq2* produced a strongly skewed MA plot when comparing PDS-treated and control LCL samples. We thus employed the loess normalization method in csaw (version 1.24.3) following differential analysis with *edgeR* (version 3.32.1) for comparing PDS- and DMSO-treated samples [81,82], as the method IV described in Yan *et al.* [83].

***Copy number normalization.*** Copy number normalization is done as described in DS et al.[48]18/09/2024 12:30:00. Briefly, prior to the differential analysis of ATAC-seq, copy number normalization was applied to the read count for peaks. The first step of copy number normalization is characterizing CNR. Genomic sequencing data from cell lines GM08505 (BS) and GM00637 (WT) were used in *CNVkit* following its recommended copy number calling pipeline *(--method wgs --target-avg-size 50000*) to identify copy number alteration in BS relative to WT [84]. Gapped regions in the human genome assembly marked by Ns (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/gap.txt.gz) were excluded from the analysis. In the output from *CNVkit*, $\log_2$-transformed values representing CNR for genomic segments were retrieved and converted back to the original value. The second step is assigning each peak to its overlapping DNA segment or the closest DNA segment. The CNR of this segment was then used as a scaling factor to modify the read/fragment count in this peak. Specifically, for the BS vs. WT comparison, if the CNR is greater than or equal to 1 (CNR ≥ 1), the fragment counts in peaks in BS were divided by the CNR. Conversely, if the CNR was smaller than 1 (CNR < 1), the fragment counts in peaks in WT were multiplied by the CNR.

**RNA-seq data analysis**

***Differential analysis of gene expression.*** RNA-seq reads from fibroblast samples were aligned to hg38 with gtf file (comprehensive gene annotation file for regions CHR from release 40 (GRCh38.p13, *https://www.gencodegenes.org/human/release_40.html*) and STAR

(version 2.7.9a) with default parameters [85]. RNA-seq reads from LCL samples were aligned to a composite reference genome of human and EBV (EBV genome is downloaded from NCBI, NC_007605.1) with the corresponding composite gtf files (gff3 for EBV is created from NC_007605.1) via *STAR* [85]. Taking advantage of the directionality of the RNA-seq library preparation protocol, the count of reads mapping to the reverse strand was used for subsequent differential analysis in *DESeq2* with default parameters.

**ChIP-seq data analysis**

***Peak calling and filtering.*** We processed the reads until the step of read filtering in the same way as described in the ATAC-seq reads processing for LCL and fibroblasts. For each biological replicate, a pooled file was generated by subsampling 35 million reads from each technical replicate. The pooled file was subjected to peak calling using *MACS2* (version 2.1.1.20160309) with *callpeak –format BAM narrowPeak* [53]*.* Peak sets from two biological replicates were ranked by their signal values and then filtered based on reproducibility using *IDR* (version 2.0.3) [54]. The final high confidence peaks for each sample were obtained with the criteria of irreproducible discovery rate (IDR) < 0.05. Peaks were also called with input samples using the same parameters, which were later used as regions to be excluded in the differential analysis.

***Differential analysis.*** The differential analysis was carried out in *DiffBind* [56]. Signals were quantified with default parameters *summits = FALSE, bFullLibSize = FALSE* followed by data normalization and statistical test in *edgeR*. We define significantly differential G4 forming sites as those with FDR < 0.05. For Fib-BS vs. Fib-WT, after signal quantification and before to data normalization, the read count was extracted from *DiffBind* and subjected to copy number normalization, as described in DS et al. [48].

**Peak annotation and functional enrichment analysis**

ATAC-seq and ChIP-seq peaks were annotated in *ChIPseeker* with the gtf file used in RNA-seq read alignment with default parameters [86]. Gene Ontology enrichment analysis on gene lists was carried out with *ChIPseeker* and *Reactome* in *R* [86,87]. Alternatively, significantly DE genes were ranked by the fold change and subjected to gene set enrichment analysis against the human hallmark gene sets with *msigdbr* in *R*.

**Peak signal profiles**

Bamfiles were converted to bigwig format via *bamCoverage* in *deeptools* with the option *--binSize 20 -of bigwig --normalizeUsing RPKM* [88]. Subsequently, bigwig files were used to compute the signal in regions of interest (TSS +/- 3kb, gene body) via *computeMatrix* function followed by function *plotHeatmap* and/or *plotProfiles* in *deeptools* with default parameters.

**Principal component analysis (PCA)**

A read count matrix with the raw count was created by summarizing the fragment count in peaks for ATAC-seq and ChIP-seq and the read count per gene for RNA-seq. The read count matrix was then transformed with the *DESeq2* function *varianceStabilizingTransformation*, following the steps outlined in the *DESeq2* manual [47]. Subsequently, PCA analysis was carried out using the *R* function *prcomp* with the option *scale = FALSE* on the processed read count matrix. The coordinate and the percentage of variance explained by each component were extracted and plotting was done with *ggplot2*.

**Converting ATAC-seq, RNA-seq and ChIP-seq signals to z-scores**

For ATAC-seq and ChIP-seq, we first converted raw fragment count in peaks to RPKM (reads per kilobase per million reads) using the total number of fragments in peaks as the total library size. For RNA-seq, we converted the fragment count per gene to TPM (transcripts per million) using the mean transcript length of each gene [49]. RPKM and TPM values are then converted to z score via *zFPKM* package in R with the function *zFPKM* [50]. Genes whose expression level has a z-score of smaller than -3 is defined as inactive or not expressed [50].

**Correlation between differential signals in RNA-seq and ATAC-seq or ChIP-seq**

In this study, the promoter is defined as the upstream 3kb (-3kb) to downstream 3kb (+3kb) region of a transcription start site (TSS). A gene will be assigned as the target gene of an ATAC-seq peak or G4 peak if the peak is less than 3kb away from the TSS. Correlation between them is assessed by comparing $log_2$FC values of RNA-seq and ChIP-seq for genes with G4 peak in their promoter in *cor.test*.

**Correlation between differential signals in ATAC-seq or ChIP-seq**

ATAC-seq and ChIP-seq peaks were intersected. An ATAC-seq peak is considered as a potential target of a G4 ChIP-seq peak if their overlap is larger than 150bp. Pearson correlation between them is tested with $log_2$FC values of ATAC-seq and ChIP-seq in *cor.test*.

**Inferring the strandedness of G4 peaks**

We downloaded the G4seq hits identified in purified human genomic DNA upon the addition of $K^+$ in the sequencing buffers via G-quadruplex induced polymerase stalling (GEO: GSE63874, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE63874, GSE63874_Na_K_minus_hits_intersect.bed.gz, GSE63874_Na_K_plus_hits_intersect.bed.gz) and lifted their coordinate to hg38 via *LiftOver* with the corresponding chain file (https://hgdownload.cse.ucsc.edu/goldenpath/hg19/liftOver/hg19ToHg38.over.chain.gz)[26,89]. A G4 peak is postulated to Watson strand and if G4 peak only intersect G4-seq hits on Watson strand. Based on the strandedness of the gene that the G4 peaks is assigned to, G4 structure forming on the transcribing or non-transcribing strand can be inferred.

**Testing interaction between G4 peaks and ATAC-seq**

We test the relationship between differential gene expression(diff-RNA), differential chromatin accessibility (diff-ATAC), and differential G4 formation (diff-G4), we used the *Car* package in *R* and built a multifactor ANOVA model with interactions and type-III test. The model was *diff-RNA = diff-ATAC + diff-G4 + diff-ATAC x diff-G4*. Values for diff-RNA is the actual $log_2$FC whereas diff-ATAC and diff-G4 are categorical values of their differential status, non-differential, up-regulated (p-adj < 0.05 or FDR < 0.05 and $log_2$FC(BS/WT) > 0), or down-regulated (up-regulated (p-adj < 0.05 or FDR < 0.05 and $log_2$FC(BS/WT) < 0). We then carried *type-III* tests, a traditional multifactor ANOVA model with interactions.

**Enrichment of G4 peaks at certain genetic features**

Enrichment is assessed by comparing the observed overlap between G4 peaks and certain genetic features to the expected number of overlaps by chance. Using TSS as an example, to evaluate if the G4 peaks are enriched at the TSS of expressed genes, we first calculated the

overlap between G4 peaks and TSS of expressed genes. Then we randomly permuted the genomic locations of G4 peaks in the genome with *bedtools shuffle* and counted the number of overlaps between randomly shuffled G4 peaks and TSS. The expected number of overlaps is represented by the mean number of overlaps in the permutation for 1000 times. We evaluated if the G4 peaks are particularly enriched at the TSS of DE genes via the function *permTest()* in *regionR* with the following parameter, *randomize.function=resampleRegions, evaluate.function=numOverlaps.* The TSS of DE genes were permutated in the TSS set of all expressed genes for 1000 times.

**Summary of statistical tests**

All statistical tests were carried out in *R*. Person correlation was tested between ATAC $log_2$FC and RNA $log_2$FC, between G4 ChIP $log_2$FC and ATAC $log_2$FC, and between G4 ChIP $log_2$FC and RNA $log_2$FC (Fig. 2a, 2b, 2d; Fig 5h; Suppl. Fig. 1b,1d; Suppl. Fig. 4a, 4b, 4d; Suppl. Fig. 5c,5e; Suppl. Fig. 7e; Suppl. Fig. 8; Suppl. Fig. 9c, 9d). Wilcoxon tests were used to compared the open chromatin accessibility, size and density of ATAC-seq peaks partitioned by G4 presence, gene expression levels partitioned by G4 presence in the promoter, gene expression levels stratified by the strandedness of G4, effect size of chromatin accessibility and G4 formation changes on gene expression changes (Fig. 5a, 5b, 5d, 5f, 5g; Fig 5h; Suppl. Fig. 7f, 7g, 7h; Suppl. Fig. 9b). Z-tests were carried out in assessing if G4 peaks or G4-seq hits are enriched in certain genomic intervals or genetic features. the permutation to compare observed values and expected values (Fig. 2f; Fig. 4f; Suppl. Fig. 4f). Hypergeometric tests (*phyper* in R) were performed to test if there is an enrichment of BS+ PDS+ and BS- PDS- signals with *lower.tail = FALSE* and if there is a depletion of BS+ PDS- and BS- PDS+ signals with *lower.tail = TRUE* (Fig. 3c, 3d).

## Supplementary information

Supplementary figures 1-9

Supplementary tables 1-4

## Acknowledgments

We are grateful to the Bloom Syndrome Registry and BS family members. We thank members in Frank Chan and Felicity Jones' lab for helpful discussions. We thank Yinan Wang, Detlef Weigel, and Marja Timmermans for their input. We also thank the Genome Center at the Max Planck Institute for Biology Tübingen for providing support. We are grateful to Linda Chen for the graphic design. We thank Marek Kučka for bioinformatic support. We thank Andre Noll for computing support.

## Funding

D.S. and M.P. are supported by an International Max Planck Research School fellowship. The research was supported by the Max Planck Society.

## Author information

**Authors and Affiliations**

**Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany**

Dingwen Su, Veronika Altmannová, Moritz Peters, Volker Soltys, John Weir, Yingguang Frank Chan

**University of Groningen, Groningen Institute of Evolutionary Life Sciences (GELIFE), 9747 AG Groningen, The Netherlands**

Yingguang Frank Chan

**Department of Pediatrics, Weill Cornell Medical College, New York, NY, USA.**

Maeve Flanagan, Christopher M Cunniff

**Corresponding author**

Correspondence to Yingguang Frank Chan (frank.chan@rug.nl)

## Ethics declarations

### Ethics approval

Not applicable.

### Competing interest

The authors declare no competing interest.

## Reference

1. Bloom, D. Congenital telangiectatic erythema resembling lupus erythematosus in dwarfs; probably a syndrome entity. *AMA Am J Dis Child* **88**, 754–758 (1954).
2. Ellis, N. A. *et al.* The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655–666 (1995).
3. Karow, J. K., Chakraverty, R. K. & Hickson, I. D. The Bloom's Syndrome Gene Product Is a 3′-5′ DNA Helicase. *Journal of Biological Chemistry* **272**, 30611–30614 (1997).
4. Cunniff, C., Bassetti, J. A. & Ellis, N. A. Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol Syndromol* **8**, 4–23 (2017).
5. Mohaghegh, P. The Bloom's and Werner's syndrome proteins are DNA structure-specific helicases. *Nucleic Acids Research* **29**, 2843–2849 (2001).
6. Hickson, I. D. RecQ helicases: caretakers of the genome. *Nat Rev Cancer* **3**, 169–178 (2003).
7. Chu, W. K. & Hickson, I. D. RecQ helicases: multifunctional genome caretakers. *Nat Rev Cancer* **9**, 644–654 (2009).
8. Chester, N., Kuo, F., Kozak, C., O'Hara, C. D. & Leder, P. Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. *Genes Dev.* **12**, 3382–3393 (1998).
9. van Wietmarschen, N. *et al.* BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat Commun* **9**, 271 (2018).
10. Chaganti, R. S. K., Schonberg, S. & German, J. A Manyfold Increase in Sister Chromatid Exchanges in Bloom's Syndrome Lymphocytes. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4508–4512 (1974).
11. Wu, L. & Hickson, I. D. The Bloom's syndrome helicase suppresses crossing over during homologous recombination. *Nature* **426**, 870–874 (2003).
12. Lazzarano, S. *et al.* Genetic mapping of species differences via in vitro crosses in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3680–3685 (2018).
13. Song, J. H. T. *et al.* Genetic studies of human–chimpanzee divergence using stem cell fusions. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2117557118 (2021).
14. Nichols, C. A. *et al.* Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. *Nat Commun* **11**, 2517 (2020).
15. Jeggo, P. A., Pearl, L. H. & Carr, A. M. DNA repair, genome stability and cancer: a historical perspective. *Nat Rev Cancer* **16**, 35–42 (2016).
16. Wu, S., Zhu, W., Thompson, P. & Hannun, Y. A. Evaluating intrinsic and non-intrinsic cancer risk factors. *Nat Commun* **9**, 3490 (2018).
17. Dutertre, S. *et al.* Cell cycle regulation of the endogenous wild type Bloom's syndrome DNA helicase. *Oncogene* **19**, 2731–2738 (2000).
18. Lee, J. *et al.* Bloom syndrome patients and mice display accelerated epigenetic aging. *Aging Cell* **22**, e13964 (2023).
19. Johnson, J. E., Cao, K., Ryvkin, P., Wang, L.-S. & Johnson, F. B. Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. *Nucleic Acids Res* **38**, 1114–1122 (2010).

20. Nguyen, G. H. *et al.* A Small Molecule Inhibitor of the BLM Helicase Modulates Chromosome Stability in Human Cells. *Chemistry & Biology* **20**, 55–62 (2013).
21. Sun, H., Karow, J. K., Hickson, I. D. & Maizels, N. The Bloom's Syndrome Helicase Unwinds G4 DNA. *Journal of Biological Chemistry* **273**, 27587–27592 (1998).
22. Hänsel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* **18**, 279–284 (2017).
23. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research* **33**, 2908–2916 (2005).
24. Puig Lombardi, E. & Londoño-Vallejo, A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Research* **48**, 1–15 (2020).
25. Hon, J., Martínek, T., Zendulka, J. & Lexa, M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics* **33**, 3373–3379 (2017).
26. Chambers, V. S. *et al.* High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol* **33**, 877–881 (2015).
27. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* **13**, 551–564 (2018).
28. Dréau, A., Venu, V., Avdievich, E., Gaspar, L. & Jones, F. C. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun* **10**, 4309 (2019).
29. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**, 1267–1272 (2016).
30. Esnault, C. *et al.* G4access identifies G-quadruplexes and their associations with open chromatin and imprinting control regions. *Nat Genet* **55**, 1359–1369 (2023).
31. Lago, S. *et al.* Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat Commun* **12**, 3885 (2021).
32. Mao, S.-Q. *et al.* DNA G-quadruplex structures mold the DNA methylome. *Nat Struct Mol Biol* **25**, 951–957 (2018).
33. Varshney, D. The regulation and functions of DNA and RNA G-quadruplexes. 16.
34. Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627–8637 (2015).
35. Lee, W. T. C. *et al.* Single-molecule imaging reveals replication fork coupled formation of G-quadruplex structures hinders local replication stress signaling. *Nat Commun* **12**, 2525 (2021).
36. Lerner, L. K. & Sale, J. E. Replication of G Quadruplex DNA. *Genes* **10**, 95 (2019).
37. Lee, C.-Y. *et al.* R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. *Nat Commun* **11**, 3392 (2020).
38. Li, C. *et al.* Ligand-induced native G-quadruplex stabilization impairs transcription initiation. *Genome Res.* **31**, 1546–1560 (2021).
39. Esain-Garcia, I. *et al.* G-quadruplex DNA structure is a positive regulator of *MYC* transcription. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2320240121 (2024).
40. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c- *MYC* transcription. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11593–11598 (2002).

41. Cogoi, S. & Xodo, L. E. G-quadruplex formation within the promoter of the KRAS proto-oncogene and its effect on transcription. *Nucleic Acids Research* **34**, 2536–2549 (2006).
42. Marquevielle, J. *et al.* Structure of two G-quadruplexes in equilibrium in the KRAS promoter. *Nucleic Acids Research* **48**, 9336–9345 (2020).
43. Fernando, H. *et al.* A conserved quadruplex motif located in a transcription activation site of the human c-kit oncogene. *Biochemistry* **45**, 7854–7860 (2006).
44. Balasubramanian, S., Hurley, L. H. & Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat Rev Drug Discov* **10**, 261–275 (2011).
45. Spiegel, J. *et al.* G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol* **22**, 117 (2021).
46. Nguyen, G. H. *et al.* Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9905–9910 (2014).
47. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
48. Su, D., Peters, M. A., Soltys, V. & Chan, Y. F. Copy number normalization distinguishes differential signals driven by copy number differences in ATAC-seq and ChIP-seq. Preprint at https://doi.org/10.1101/2024.04.11.588815 (2024).
49. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
50. Hart, T., Komori, H., LaMere, S., Podshivalova, K. & Salomon, D. R. Finding the active genes in deep RNA-seq gene expression studies. *BMC Genomics* **14**, 778 (2013).
51. Grierson, P. M. *et al.* BLM helicase facilitates RNA polymerase I-mediated ribosomal RNA transcription. 12.
52. Schawalder, J., Paric, E. & Neff, N. F. Telomere and ribosomal DNA repeats are chromosomal targets of the bloom syndrome DNA helicase. *BMC Cell Biology* 14 (2003).
53. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
54. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, (2011).
55. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME Suite. *Nucleic Acids Res* **43**, W39–W49 (2015).
56. Stark, R. & Brown, G. DiffBind: Differential binding analysis of ChIP-Seq peak data.
57. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).
58. Lobo, T. J., Lansdorp, P. M. & Guryev, V. Local G-quadruplexes are not a major determinant of altered gene expression in BLM-deficient cells. Preprint at https://doi.org/10.1101/2023.09.08.556664 (2023).
59. Manthei, K. A. & Keck, J. L. The BLM dissolvasome in DNA replication and repair. *Cell. Mol. Life Sci.* **70**, 4067–4084 (2013).
60. Moruno-Manchon, J. F. *et al.* Small-molecule G-quadruplex stabilizers reveal a novel pathway of autophagy regulation in neurons. *eLife* **9**, e52283 (2020).
61. Li, J.-L. *et al.* Inhibition of the Bloom's and Werner's Syndrome Helicases by G-Quadruplex Interacting Ligands. *Biochemistry* **40**, 15194–15202 (2001).

62. German, J., Sanz, M. M., Ciocci, S., Ye, T. Z. & Ellis, N. A. Syndrome-causing mutations of the BLM gene in persons in the Bloom's Syndrome Registry. *HUMAN MUTATION* 12 (2007).
63. King, J. J. *et al.* Metastable Intermediates Identified in Epithelial to Mesenchymal Transition are Regulated by G-Quadruplex DNA Structures. Preprint at https://doi.org/10.1101/2023.08.21.554220 (2023).
64. Tangeman, L., McIlhatton, M. A., Grierson, P., Groden, J. & Acharya, S. Regulation of BLM Nucleolar Localization. *Genes (Basel)* **7**, 69 (2016).
65. Thiery, J. P., Acloque, H., Huang, R. Y. J. & Nieto, M. A. Epithelial-mesenchymal transitions in development and disease. *Cell* **139**, 871–890 (2009).
66. Usman, S. *et al.* Vimentin Is at the Heart of Epithelial Mesenchymal Transition (EMT) Mediated Metastasis. *Cancers* **13**, 4985 (2021).
67. Ceschi, S. *et al.* Vimentin binds to G-quadruplex repeats found at telomeres and gene promoters. *Nucleic Acids Research* **50**, 1370–1381 (2022).
68. Zyner, K. G. *et al.* Genetic interactions of G-quadruplexes in humans. *eLife* **8**, e46793 (2019).
69. Mishra, S. K., Tawani, A., Mishra, A. & Kumar, A. G4IPDB: A database for G-quadruplex structure forming nucleic acid interacting proteins. *Sci Rep* **6**, 38144 (2016).
70. Lyu, J., Shao, R., Kwong Yung, P. Y. & Elsässer, S. J. Genome-wide mapping of G-quadruplex structures with CUT&Tag. *Nucleic Acids Research* **50**, e13–e13 (2022).
71. Hui, W. W. I., Simeone, A., Zyner, K. G., Tannahill, D. & Balasubramanian, S. Single-cell mapping of DNA G-quadruplex structures in human cancer cells. *Sci Rep* **11**, 23641 (2021).
72. Sauer, M. & Paeschke, K. G-quadruplex unwinding helicases and their function in vivo. *Biochemical Society Transactions* 10 (2017).
73. Sarkies, P. *et al.* FANCJ coordinates two pathways that maintain epigenetic stability at G-quadruplex DNA. *Nucleic Acids Research* **40**, 1485–1498 (2012).
74. Croteau, D. L., Popuri, V., Opresko, P. L. & Bohr, V. A. Human RecQ Helicases in DNA Repair, Recombination, and Replication. *Annu. Rev. Biochem.* **83**, 519–552 (2014).
75. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. 15 (2017).
76. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
77. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011).
78. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at http://arxiv.org/abs/1303.3997 (2013).
79. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
80. Zhu, A., Ibrahim, J. G. & Love, M. I. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics* **35**, 2084–2092 (2019).
81. Lun, A. T. L. & Smyth, G. K. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research* **44**, e45–e45 (2016).

82. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
83. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* **21**, 22 (2020).
84. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873 (2016).
85. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
86. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
87. Yu, G. & He, Q.-Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* **12**, 477–479 (2016).
88. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**, W187–W191 (2014).
89. Bioconductor Package. liftOver. https://doi.org/10.18129/B9.BIOC.LIFTOVER.
90. Fox, J., & Weisberg, S. (2019). *An R Companion to Applied Regression* (3rd ed.). Sage, Thousand Oaks, CA.

# Chapter 2

## Copy number normalization distinguishes differential signals driven by copy number differences in ATAC-seq and ChIP-seq

Dingwen Su [1*], Moritz Peters [1], Volker Soltys [1], Yingguang Frank Chan [1,2*]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany

2 University of Groningen, Groningen Institute for Evolutionary Life Sciences (GELIFES), 9747 AG Groningen, The Netherlands

* Corresponding authors

### Status in publication process

Available as preprint, https://doi.org/10.1101/2024.04.11.58881; under review at *BMC Genomics*

### Author contributions

D.S. conceived the idea, designed the experiment, generated the data, conducted the data analysis and developed the pipeline. Y.F.C. supervised the data collection and helped interpret the results. M.P., V.S. and Y.F.C. provided experimental or computational support. D.S. drafted the manuscript with input from all authors. All authors reviewed and approved the final version of the manuscript.

## Abstract

A common objective across ATAC-seq and ChIP-seq analyses is to identify differential signals across contrasted conditions. However, in differential analyses, the impact of copy number variation is often overlooked. Here, we demonstrated copy number differences among samples could drive, if not dominate, differential signals. To address this, we propose a pipeline featuring copy number normalization. By comparing the averaged signal per gene copy, it effectively segregates differential signals driven by copy number from other factors. Further applying it to Down syndrome unveiled distinct dosage-dependent and -independent changes on chromosome 21. Thus, we recommend copy number normalization as a general approach.

## Background

Recent advances in next-generation sequencing assays have boosted the application of high-throughput methods in research. Combined with rapidly decreasing costs, researchers can now routinely apply functional genomic assays such as Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq) and chromatin immunoprecipitation followed by sequencing (ChIP-seq) to profile alterations in histone modifications, protein-DNA interactions and open chromatin landscapes, thus providing information on gene regulation [1–5]. Changes in them are closely associated with early developmental processes, the initiation, and progression of diseases. In the case of cancer, differential signals can serve as potential biomarkers and even therapeutic targets [6,7].

A critical step in these studies is to identify differential signals between perturbed and control states, along developmental stages or in response to the titration of specific chemical treatments. After read alignment and filtering, a common workflow of the differential analysis for ATAC-seq and ChIP-seq is to first identify a focal set of regions with enriched signal (peaks), then quantify the signals, normalize them and lastly detect differential signals via statistical tests. It is important to note that the differential results can differ greatly depending on bioinformatic pipelines (e.g., ENCODE

pipelines) and specific parameter choices, even from the same set of data, as demonstrated by Reske *et al.* [8–12]. Even when using the same data and pipeline, the outcome of the differential analysis may differ depending on how the variation in the data is attributed.

Variations in signals between contrasted conditions arise from various sources, which is ideally but often not limited to the factor(s) of interest. Correcting biases and attributing the variation to factors accordingly in the differential analysis is crucial for obtaining an accurate set of differential signals. One factor often overlooked in the differential analysis is copy number variation (CNV), including chromosomal aberrations such as aneuploidy. Specifically, since the signal captured by ATAC-seq or ChIP-seq is the sum of signals from all gene copies, variation in the underlying copy number or ploidy can directly impact the aggregated signal and the interpretation of differential signals. As a result, differences in copy numbers between samples may drive the identification of differential signals, even without any changes in local differences in chromatin accessibility or binding.

However, commonly used tools to quantify reads/fragments in ATAC-seq or ChIP-seq peaks like *bedtools*, *deeptools*, *htseq-count,* and *featureCounts* in the first place are not inherently designed to distinguish between background and effective signals. As a result, these tools quantify background signals as effective signals, inflating the detected effective signals [13–16]. A higher copy number in a region will thus inflate the signal to a larger extent, increasing the likelihood of misidentifying this region as exhibiting elevated signals. Other tools like *DiffBind* and *csaw* do offer the option to subtract the signals from the background when quantifying the signals in peaks [17–19]. However, by being copy-number blind, the effects of differences in the copy number between samples remain. Meanwhile, CNV and/or aneuploidy are commonplace in cell lines used in biomedical research. It could arise from cancer, defective developmental defects (e.g., trisomy 13 or 21, the latter commonly known as Down syndrome), the process of establishing cell lines and repeated passages in tissue culture [20–23]. Additionally, many tissues are naturally polyploid, such as the salivary gland in *Drosophila melanogaster*, the liver in humans and the developing root in *Arabidopsis thaliana.* Therefore, to accurately identify the alterations due to the

factor(s) of interest in biological processes or diseases, it is necessary to take CNV into account as a source of variation in the differential analysis, especially when CNV and aneuploidy may arise as an artifact of cell immortalization.

To address this, we propose a differential analysis pipeline featuring copy number (CN) normalization and we showcase its application and advantages using two examples in biomedical studies. In the first case, we applied it to ATAC-seq and ChIP-seq data generated from cell lines with complex chromosomal aberrations derived from a Bloom syndrome individual and a healthy donor. Using a conventional copy-number blind pipeline, we noticed heavily skewed differential signals toward the sample with relatively higher copy numbers in the corresponding regions. In this case where CNV is not a central factor of interest, applying our pipeline with CN normalization efficiently distinguishes differential signals driven by copy number variations and those due to the disease. In the second case, we applied it to ATAC-seq data generated from trisomy 21 and euploid cell lines. By combining the results from our pipeline with the common workflow, we were able to distinguish the open chromatin regions on chromosome 21 with dosage effects, compensatory effects and copy number-independent regulatory changes.

## Results

### Copy number variation drives the identification of differential signals in ATAC-seq and ChIP-seq

Bloom syndrome (BS) is a rare autosomal recessive disorder with complex clinical manifestations. It is caused by loss-of-function of mutations in the conserved Bloom syndrome RecQ like helicase, *BLM* gene, on chromosome 15. BLM helicase untangles various aberrant DNA structures and thus plays an essential role in maintaining genome integrity [24,25]. Clinically, BS individuals have strikingly shorter lifespans, in average 26 years. They have an elevated risk of cancer and are predisposed to developing cancer at an early age [26]. At the molecular level, BS cells display characteristic excessive sister chromatid exchange events and hallmarks of genome

instability [26–28]. Identifying molecular alterations in BS helps to better understand the pathogenesis of the disease as well as the function of BLM helicase.

As primary samples from BS individuals are not readily available, we used two fibroblast cell lines derived from a BS individual and a healthy donor (wildtype or WT) and profiled their open chromatin landscape with ATAC-seq and endogenously formed G-quadruplex structures via ChIP-seq [29]. The Coriell repository described these cell lines as displaying different and atypical karyotypes as well as chromosomal rearrangement events. To assess copy number differences between the BS and WT cell lines, we used *CNVkit* and whole-genome sequencing data to estimate the copy number ratio (CNR), the ratio of local copy number in a given region in BS against WT [30]. Besides irregular karyotypes with complex aneuploidy in each cell line, there were widespread copy number differences between the two cell lines (**Fig. 1a**, **Additional file 1**). Excluding sex chromosomes and masked regions in the human reference genome (hg38), the BS sample exhibits relative copy number gains and losses in a total of 1273.1 Mb (47.0%) vs. 1437.9 Mb (53.0%), respectively.

To characterize the impacts of BS on chromatin accessibility, we performed differential analysis using a workflow with a few widely used bioinformatic tools. The major steps after read alignment and filtering included: 1) calling peaks with *MACS2*; 2) quantifying signal as the number of reads/fragments in peaks with *htseq-count*; 3) performing data normalization and identifying differential signals with *DESeq2*, specifying genotype ((BS vs. WT) as the contrasting factor [31]. Following this pipeline, we identified 143,460 accessible chromatin regions in WT and BS cells and detected 89,516 (62.40%) significantly differential peaks (herein defined as peaks with adjusted $P < 0.05$) in BS vs. WT. Under this standard pipeline, 42,831 (29.86%) peaks showed increased accessibility, while 46,685 (32.54%) displayed decreased accessibility. When assessing the genome-wide differential signals using an MA plot, we observed no obvious skewness towards the WT or BS sample (**Fig. 1b**). Crucially, we noted a trended CNR-dependent bias in the differential signals (**Fig. 1c**, **left**): there was an over-representation of accessible peaks in regions with copy number gains ($\log_2$ CNR $> 0$); conversely, less accessible peaks in regions with copy number losses ($\log_2$ CNR $< 0$).

To better investigate this effect, we examined two chromosomes in detail to see how copy number differences may affect the differential analysis. For chromosome 17 the entire chromosome showed relatively higher copy numbers in BS (**Fig. 1d**, **top**). About 72% (5239 out of 7231) of the peaks were significantly differential. Notably, there is a very strong directionality to this set: 62.40% (4486) showed increased chromatin accessibility in BS, but only 10.47% (753) with decreased signals (**Fig. 1d**, **middle**). Both of these proportions deviated strongly from the genome-wide trends of approximately 30% in either direction (Chi-squared test, $\chi^2 = 3529$, df = 2, $P < 2.2 \times 10^{-16}$). For chromosome 20, there was relative copy number loss for the BS sample in the short arm and relative copy number gain in the long arm (**Fig. 1e**, **top**). Consistently, we observed similar skewness of differential signals toward the sample with a relatively higher copy number: an overrepresentation of less accessible peaks for BS on the short arm (629 out of 874 peaks) and more accessible peaks on the long arm (1933 out of 3034 peaks) (**Fig. 1e**, **top and middle**). These observations strongly suggest that CNV was the main driver of the differential signals at these regions and may thus substantially confound the interpretation of the results.

ChIP-seq is another common quantitative genomic sequencing assay whose readout, like ATAC-seq, is based on quantifying the number of reads/fragments from a genomic interval. Here, we performed G4 ChIP-seq using antibodies against a secondary DNA structure, G-quadruplexes (G4), and mapped the endogenously formed G-quadruplexes in the aforementioned WT and BS cell lines. G-quadruplexes are a four-stranded secondary DNA structure formed by G-rich sequences through cyclic Hoogsteen hydrogen bonds and are one of the DNA substrates of BLM helicase [32,33]. In total, we identified 20,072 high-confidence G4 peaks and performed the differential analysis to characterize differential G4 forming sites via *DiffBind*. We detected 11,730 (58.5%) significantly differential sites (FDR < 0.05), of which 6217 and 5513 showed increased and decreased signals, respectively. Similar to our previous observation for ATAC-seq, we observed an even stronger CN-dependent bias in the differential signals (**Additional file 2: Fig. S1a**). Peaks identified to have decreased G4 forming signals were overrepresented in regions with relatively lower copy numbers in BS vs. WT ($\log_2$ CNR < 0) and vice versa.

Our results from both ATAC-seq and ChIP-seq showed that copy number differences between contrasted samples could drive the differential signals. Notably, while BS cells have been reported to have excessive sister chromatid exchange (SCE) events, tissues/cells directly sampled from BS individuals mostly showed normal karyotypes [26,27,34]. Occasional chromosome losses has been reported in some studies though, rarely has BS been directly linked to complex aneuploidy and broad genome-wide chromosomal aberrations in normal tissues, as observed in our cell lines [26,34–36]. Moreover, the fibroblast cell line derived from a healthy donor showed widespread chromosomal aberrations. Therefore, we interpret the chromosomal aberrations in the fibroblast cell lines to be an artifact during the immortalization of the cell line rather than a primary effect of BS. Misattributing the variation driven by CNV between WT and BS samples as differential signals due to BS would obscure the biological significance of the observed differences.

**Copy number normalization separates differential signals driven by copy number variation**

We have demonstrated the impacts of copy number variation on differential analysis. If left uncorrected, CNV could mask biologically relevant differential signals and confound the interpretation of the results. It is therefore necessary to separate the differential peaks driven by CNV and/or aneuploidy from other factors, particularly, the factor of interest, here BS. This is also of general interest because aneuploidy and CNV are common to various biological materials, e.g., cancer tissues. To account for copy number in the differential analysis, we implemented CN normalization after signal quantification, but before data normalization in the pipeline (**Fig. 2**).

Copy number normalization mainly consists of two steps, estimating and then correcting differences in copy number. The first step is to characterize the copy number ratio, which is the regional relative copy number variation in the perturbed sample relative to the control sample. Notably, we aim to detect CNR instead of an absolute genome-wide CN determination within each sample, because it does not require prior knowledge of the karyotype in any of the contrasted samples. Here, we used *CNVkit* to estimate local CNR [30]. It uses genomic sequencing data from ATAC-

seq (or ChIP-seq input) from both BS and WT samples to first calculate the coverage in non-overlapping 50kb bins in the genome, then segment the genome, and lastly assess the CNR for each segment. Next, we assigned each peak to its overlapping DNA segment or the closest DNA segment and corrected the read/fragment count in this peak by applying the CNR of the DNA segment as a scaling factor. The modified count matrix is used in the subsequent steps of data normalization and calling of differential signals (**Fig. 2**). Specifically, for peaks in regions with copy number gain ($\log_2$ CNR > 0), the read/fragment counts in BS samples were divided by CNR whereas for peaks in regions with copy number loss ($\log_2$ CNR < 0), the read/fragment counts in WT samples were multiplied by CNR to avoid inflating the statistical power to detect the differential signals. Notably, these add-on steps can be easily integrated into most pipelines using a count-based approach for signal quantification.

After having applied CN normalization, we re-analyzed the ATAC-seq data between BS and WT samples and obtained strikingly different results (**Fig. 3a, left**). Importantly, the copy number-dependent bias of differential signals was largely removed. This effectively eliminated the overrepresentation of differential signals towards the sample with a relatively higher copy number in the region (**Fig. 1c**, **right**; **Fig. 1e**, **bottom**). For instance, on chromosome 17 with relative copy number gain, fewer peaks remained significantly differential following CN normalization (3559 vs. 5238 without CN normalization) and differential peaks displayed a more balanced distribution (1822 increased and 1737 decreased) along the chromosome (**Fig. 1d**, **bottom**; **Fig. 3a, middle**). Similarly, on the long arm of chromosome 20, the overrepresentation of increased signals is attenuated (25.18% vs. 63.7% without CN normalization) (**Fig. 3a, right**). Out of 1348 peaks on the short arm with copy number loss, the significantly differential signals are comparable before and after CN normalization (775 vs. 825) (**Fig. 1e**, **bottom**; **Fig. 3a, right**). We also noted that on the short arm, a large majority of the significantly less accessible regions retained their differential status after CN normalization, suggestive of intrinsic regulatory changes in these regions in BS. Alternatively, this could also be attributed to relatively small copy number differences in these regions, resulting in less CNV-driving bias in the differential signals.

To characterize the global impact of CN normalization on the outcome of differential analysis, we determined how the differential status has changed for each open chromatin region in ATAC-seq data or G4 peak in ChIP-seq data following CN normalization. Overall, about 20% of the ATAC-seq peaks changed their differential status, with most of these now re-classified as non-differential (adjusted $P \geq 0.05$; **Fig. 3a, left**; **Fig. 3b**). Notably, the CN normalization has remarkably stronger impacts on the ChIP-seq data. The number of significantly differential G4 forming sites dropped by 68% to only 3704 from 11730 without CN normalization (**Fig. 3c**). These observations suggest that differential signals identified with the usual workflow are substantially confounded by copy number variation, particularly in ChIP-seq data. Crucially, we noted that our ChIP-seq data had a lower signal-to-noise ratio compared to the ATAC-seq data (**Additional file 2: Fig. S1b**). This has implications for signal quantification. This is because during counting reads/fragments at peaks, the increased background signal could also be counted as signals. Importantly, this contribution from background reads increased in proportion to the copy number. As a result, data with worse signal-to-noise ratios would be biased to a larger extent by the differences in copy numbers between samples. Consistent with this effect, peaks in regions with extreme CNR were prone to be identified as differential upon no CN normalization.

## Copy number normalization identifies accessible chromatin regions with dosage effects and compensatory effects in Down syndrome

So far, we have examined a case where CN differences is not thought to be the primary disease-causing mechanism. Next, we investigated the impact of CN differences in Down syndrome (DS), a classical example of aneuploidy. Down syndrome, also known as trisomy 21, is a genetic disorder caused by harboring an extra copy (95% of the cases) of chromosome 21. It is the most common chromosomal anomaly and the most common genetic cause of significant intellectual disability [37,38]. We used publicly available ATAC-seq data from paired fraternal lymphoblast cell lines from DS (47, XY, +21) and non-DS (46, XY) individuals (referred to as wildtype or WT in the subsequent text) [39]. In total, we identified 1397 open chromatin regions on chromosome 21. No open chromatin region was in the 34 kb highly restricted Down

syndrome critical region (hg38, chr21: 37,929,229 - 37,963,130) (**Additional file 2: Fig. S3**), the minimal region whose triplication is shared by all DS subjects and is absent in all non-DS subjects [40]. In general, open chromatin regions on chromosome 21 displayed higher chromatin accessibility in the DS sample compared to the euploid WT sample. By contrast, on chromosome 17, where both DS and WT samples are diploid, we observed comparable signals between WT and DS samples (**Fig. 4a**, **4b**). This observation is consistent with our observation with BS vs. WT samples that elevated chromatin accessibilities were associated with a relatively higher copy number.

Here, aneuploidy is the central cause of DS: the differential signals due to CNV reflect the direct impact of copy number gain of chromosome 21 and are thus biologically relevant. Therefore, we performed the genome-wide differential analyses with and without CN normalization. They produced strikingly different results for peaks on chromosome 21. Without applying CN normalization, trisomy 21 overwhelmingly drives the differential accessibility signal on chromosome 21. Out of the 457 significantly differential accessible regions, 430 (94.1%) regions showed increased chromatin accessibility while only 27 regions with decreased accessibility (**Fig. 4c**, **top**). With CN normalization, we identified only 165 significantly differential signals with 69 displaying increased signals and 96 showing decreased signals, including the same 27 regions that exhibited decreased signals without CN normalization. (**Fig. 4c**, **bottom**).

In the context of DS, differential signals identified by these two approaches call for distinct biological interpretations. One direct impact of trisomy 21 is the gene dosage and potential dosage effects which are determined by the amount of expressed gene product. There have been extensive studies in DS reporting increased transcription and protein expression levels of genes on chromosome 21, demonstrating gene dosage effects [41–46]. Meanwhile, in many other biological contexts, mechanisms engage to prevent expression levels from increasing linearly with CN, e.g., dosage compensation [41,45,47,48]. Here, we should look beyond expressed gene product, and extend these concepts to regulatory signals of chromatin such as chromatin accessibility and chromatin binding. Using ATAC-seq data and chromosome 21 in DS

as an example, the standard differential analysis pipeline compares the total signals from three gene copies in the DS sample with those from two gene copies in the WT sample (**Fig. 4d**, **left**). This approach reveals differential signals linked to the presence of an extra chromosome 21. For instance, regions with increased signals imply dosage effects of chromatin accessibility. In contrast, within the CN-aware pipeline incorporating CN normalization, the average signal per-chromosome copy is compared. Specifically, the total signal from all three copies in DS samples is divided by the copy number ratio (CNR = 3/2). Subsequently, the aggregated chromatin accessibility adjusted for two gene copies is compared against the signal from WT to reveal CN-independent changes (**Fig. 4d**, **left**). For instance, regions with decreased signals suggest dosage compensation of chromatin accessibility in response to trisomy 21.

As discussed above, these two approaches identify CN-dependent and -independent differential signals in DS. Based on the combination of the changes in total and per-chromosome copy signals, peaks on chromosome 21 could be divided into the following categories: I) regions with no change; II) regions with over-compensatory effects and CN-independent decreases, showing decreased total and per-copy signal; III) regions with compensatory effects, in which total signals remain unchanged but per-copy signal decreases; IV) regions showing only dosage effects, with increased total signals but no change in per-copy signal; V) regions with both dosage effects and CN-independent increases (**Fig. 4d**, **right**). Out of 940 regions showing no change in total chromatin accessibility in DS (category I and II), the conventional pipeline would misidentify all of them as regions with compensatory effects, whereas CN normalization enabled us to show that only 69 (7.3%) have compensatory effects (**Fig. 4d**, **right**). Moreover, among all the regions with increased total chromatin accessibility in DS (category IV and V), CN normalization distinguished that approximately 80% are purely due to dosage effects, with the rest exhibiting further CN-independent increases (**Fig. 4d**, **right**).

Many of the identified differential signals are in genes associated with the clinical symptoms of DS. It has been reported that DS individuals are more susceptible to Alzheimer's disease as the gene encoding amyloid precursor protein, *APP*, is located

on chromosome 21 [37,38]. The peak in the promoter (in this study defined as TSS ± 3kb) of *APP* showed only dosage-dependent increases whereas another peak in the promoter of a splicing variant showed a further increase in the per-copy signal, which may further boost the expression of *APP* and that specific transcript variant (**Fig. 4e**, **top**). Additionally, DS individuals exhibit higher risks of autoimmune diseases and developing severe responses to infectious diseases [37,38,49–51]. Tissues and cells from DS individuals and mouse models suggest that DS exhibits interferon hyperactivity and chronic inflammation [42,52,53]. One of the major reasons for dysregulated immune function in DS is that chromosome 21 carries genes essential for immune function. A T-cell ubiquitin ligand coding gene on chromosome 21, SH3 domain-containing protein A (*UBASH3A*), negatively regulates T-cell signaling and it is a risk gene for autoimmune diseases [54]. Almost all the open chromatin regions in this gene displayed a decreased average signal on each chromosome 21 such that the total chromatin accessibility remained the same in DS, suggesting the critical functions and tight regulatory mechanisms associated with this gene (**Fig. 4e**, **bottom**). Moreover, four of the six interferon receptor genes are also on chromosome 21. Among the 24 peaks in this interferon receptor gene cluster, 14 showed increased chromatin accessibility due to dosage effects, and one peak in the intronic regions of *IFNAR1* and two peaks in the intergenic regions showed CN-independent changes (**Additional file 2: Fig. S4a**). Another gene on chromosome 21 that is associated with inflammatory signals in the brain in DS is *S100B* [38]. Despite our data generated from lymphocytes rather than from brain tissues, ATAC-seq peaks near this gene showed dosage effects and increased chromatin accessibility per copy (**Additional file 2: Fig. S4b**). The above examples illustrate how combining overall chromatin accessibility with per-allele regulatory changes can further help explain DS phenotypes apart from transcription dosage effects. These CN-independent regulatory changes would have been missed without incorporating the CN normalization step.

As further evidence on the robustness of the procedure, we noted that the vast majority of differential signals on the other chromosomes except chromosome 21 remained the same (**Additional file 2: Fig. S2b, 2c and 2d**), suggesting that the current pipeline is

specific in its ability to correct for effect due solely to local copy number differences between contrasted samples.

## Discussion

Here we have demonstrated that copy number variation among contrasted samples, a factor often overlooked in the differential analysis, can and does drive much of the observed differential signal. Given the prevalence of aneuploidy in diseases such as cancer and cell lines used in biomedical research, such effects could hold clinical relevance. To address this, we developed a copy-number-aware pipeline featuring copy number normalization, which accounts for the copy number differences between samples and adjusts the signals to the same copy number before downstream data normalization. We presented two cases in biomedical studies to illustrate the impacts of copy number differences on the differential analysis. We further demonstrated that applying our proposed CN-aware pipeline can effectively segregate the CN-dependent from the CN-independent differential signals.

In the case of Bloom syndrome, caused by loss-of-function mutations in *BLM*, cell lines derived from both the healthy donor and the BS individual accumulated complex chromosomal aberrations irrelevant to the disease status. We observed that without accounting for CNV between the samples, differential signals are biased toward the sample with a higher copy number. The copy number differences affected approximately 20% of differential signals in ATAC-seq and 70% in ChIP-seq data. These false-positive signals would be attributed to BS, leading to potential misinterpretation of results. By applying CN normalization, we largely removed the CN-dependent differential signals. In the case of Down syndrome, trisomy 21 is the central cause of the disorder and thus the main factor of interest. By performing differential analyses with and without CN normalization, we were able to detect CN-dependent and -independent changes. Combining results from both approaches enables a nuanced biological interpretation of the changes in chromatin accessibility. By extending the concepts of dosage effects and dosage compensation to chromatin accessibility, we revealed the subset of peaks with genuine dosage compensatory

effects and distinguished peaks with only dosage effects from those with further CN-dependent increases.

Such regions with CN-independent changes would have been overlooked without considering CNV, yet they are potentially clinically relevant. For instance, most of the peaks in the risk gene for Alzheimer's disease showed increased chromatin accessibility mainly due to the extra copy of chromosome 21. One peak in the promoter region of a transcript variant of *APP* gene exhibited further CN-independent increases, potentially boosting the expression of a specific variant of the APP protein and contributing to an increased risk of Alzheimer's disease in individuals with DS. Additionally, most of the peaks in the overexpressed interferon receptor genes showed pure dosage effects [42,49,52]. Conversely, in an autoimmune-related gene, *UBASH3A*, nearly all the peaks in this gene showed decreased chromatin accessibility, possibly in compensation for the third gene copy, implying a critical role and tight regulation of this gene. These CN-independent changes plausibly arise secondarily from the primary molecular changes upon trisomy 21 and provide novel insights into dosage imbalance and the pathogenesis of DS. Further studies on these regions are required to understand the disrupted epigenetic regulatory network. In particular, performing single-cell assays on DS samples to tangle the activation and silencing of different chromosome 21 copies will provide deeper insights into dosage and compensatory effects of chromatin accessibility and gene products, which help understand the changes in regulatory network in DS.

Notably, instead of determining the exact CNV within each sample, our CN normalization aims to characterize and utilize the regional copy number ratio between the contrasted samples. Detecting the former can be challenging as it requires prior knowledge of the karyotypes, which may not be readily available, especially for samples with complex CNV and aneuploidy, such as cancer cells. In contrast, characterizing the CNR circumvents this issue, although there are scenarios where direct per-sample CNV determination may be preferred. Here we used *CNVkit* with a bin size of 50kb to estimate CNR [30]. This tool estimates the copy number by comparing the coverage of samples to a reference sample in each bin and then segments the genome based on variation in the coverage. The resolution of CNR can be improved

by deeper sequencing and using smaller bin sizes. Alternatively, *CNVkit* can also call the exact CNV within each individual sample. Additionally, *CNVkit* offers a convenient all-in-one command. Nevertheless, other tools such as *CopywriteR*, *QDNAseq* and pipelines like the Genome Analysis Toolkit (*GATK*) best practice for Somatic copy number variant discovery can also be employed to characterize CNV and calculate CNR [55–59]. Once CNR is obtained, assigning CNR as scaling factors to peaks and adjusting the quantified signals do not rely on any specific bioinformatic tools. These steps are not exclusive to our workflow or our choices of bioinformatic tools; instead, they can be implemented independently as add-on steps to most differential analysis pipelines for ATAC-seq and ChIP-seq.

Our pipeline mainly focuses on addressing the impacts of CNV in differential analysis and not in the preceding step of identifying regions with enriched signals or peak calling. Incorporating tools like Histone Modification in Cancer (*HMCan*), which uses Hidden Markov Model (HMM) and accounts for copy number variation to call peaks, could be integrated into our proposed pipeline to eliminate false peaks and further improve the results [60]. The same group also developed *HMCan-diff* for differential analysis counting for copy number variation [61]. While we agree with taking CNV into account, this tool differs from our recommendation in that it aims to be a stand-alone pipeline. As such, this tool may limit the flexibility of using other statistical models in data normalization methods and testing for differential signals.

Apart from ATAC-seq and ChIP-seq, variation in copy number, being the baseline of DNA substrate, will also be a significant factor in the differential analysis of other functional genomic assays that employ count-based approaches to quantify signals from DNA and/or chromatin, like DNase I hypersensitive sites sequencing (DNase-seq), micrococcal nuclease digestion with deep sequencing (MNase-seq), Cleavage Under Targets and Release Using Nuclease (CUT&RUN), Cleavage Under Targets and Tagmentation or (CUT&Tag). Notably, our CN normalization strategy is not limited to ATAC-seq and ChIP-seq but can also be applied to the aforementioned data types. In our implementation, the CN normalization steps can be seamlessly integrated into the differential analysis pipelines of these assays as add-on steps.

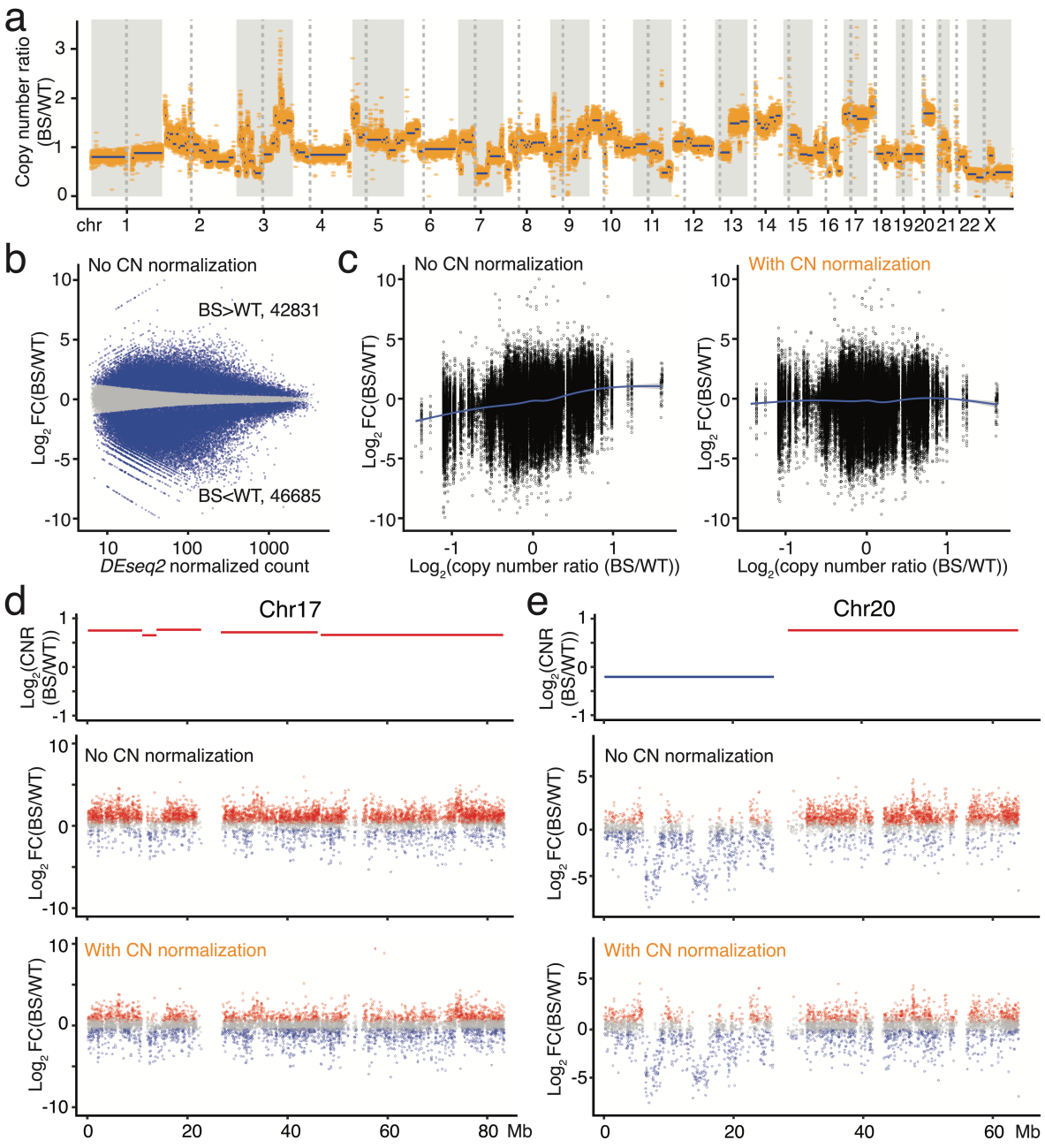



**Figure 1. Copy number variation drives the differential signals in ATAC-seq data.** (**a**)Genome-wide copy number ratio (CNR) in the BS sample relative to the WT sample. Orange dots indicate the CNR for each 50kb bin and blue lines indicate DNA segments with the same CNR. (**b**) MA plot comparing the chromatin accessibility in BS vs. WT samples by *DESeq2*. FC, fold change. (**c**) The trended biases of differential chromatin accessibility from copy number differences without (left) and with (right) copy number (CN) normalization in ATAC-seq data. $Log_2CNR > 0$ and $log_2CNR < 0$ imply relative number gain and loss in BS, respectively. The lines represent the locally weighted running line smoother (LOESS) smoothing curve for the data with the grey area indicating the 95% confidence interval for the fitted curve. (**d**) CNR and the distribution of differential peaks without and with CN normalization on chromosome 17

(chr17). Peaks with adjusted *P* (p-adj) < 0.05 are depicted in blue or red while those with p-adj ≥ 0.05 are shown in grey. (**e**) CNR and the distribution of differential peaks without and with CN normalization on chr20. Peaks with p-adj < 0.05 are depicted in blue or red while those with p-adj ≥ 0.05 are shown in grey.

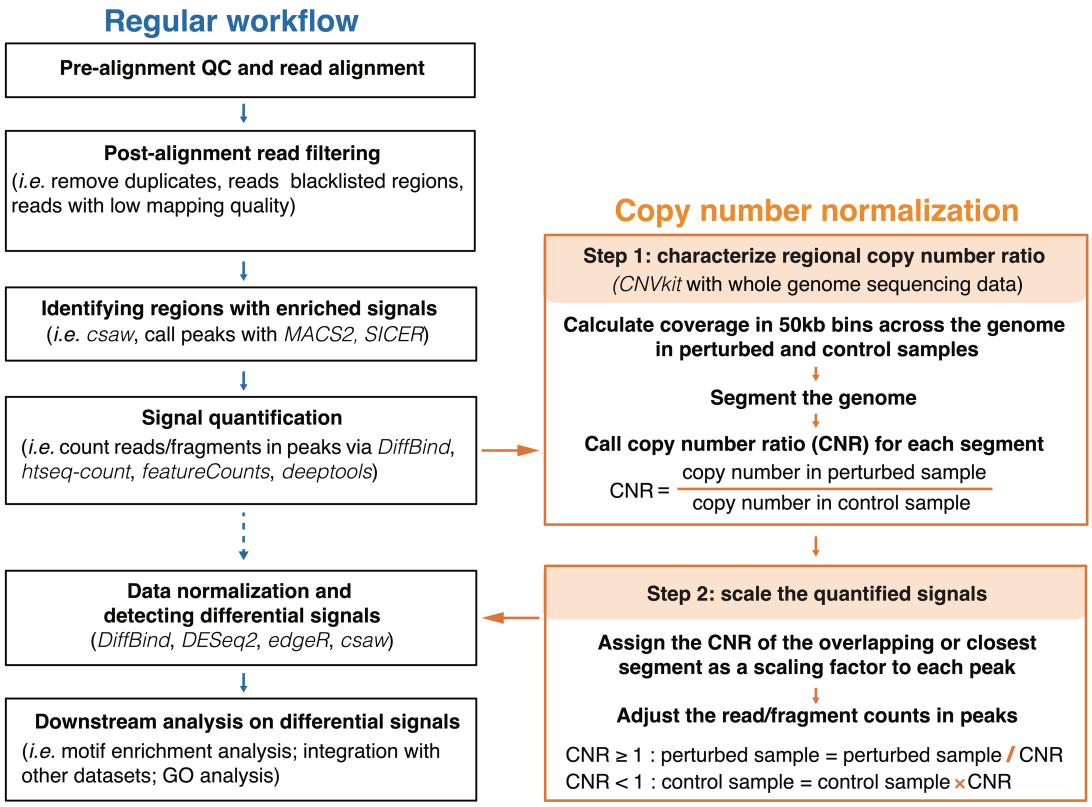



**Figure 2. A copy-number-aware differential analysis pipeline featuring copy number normalization.**

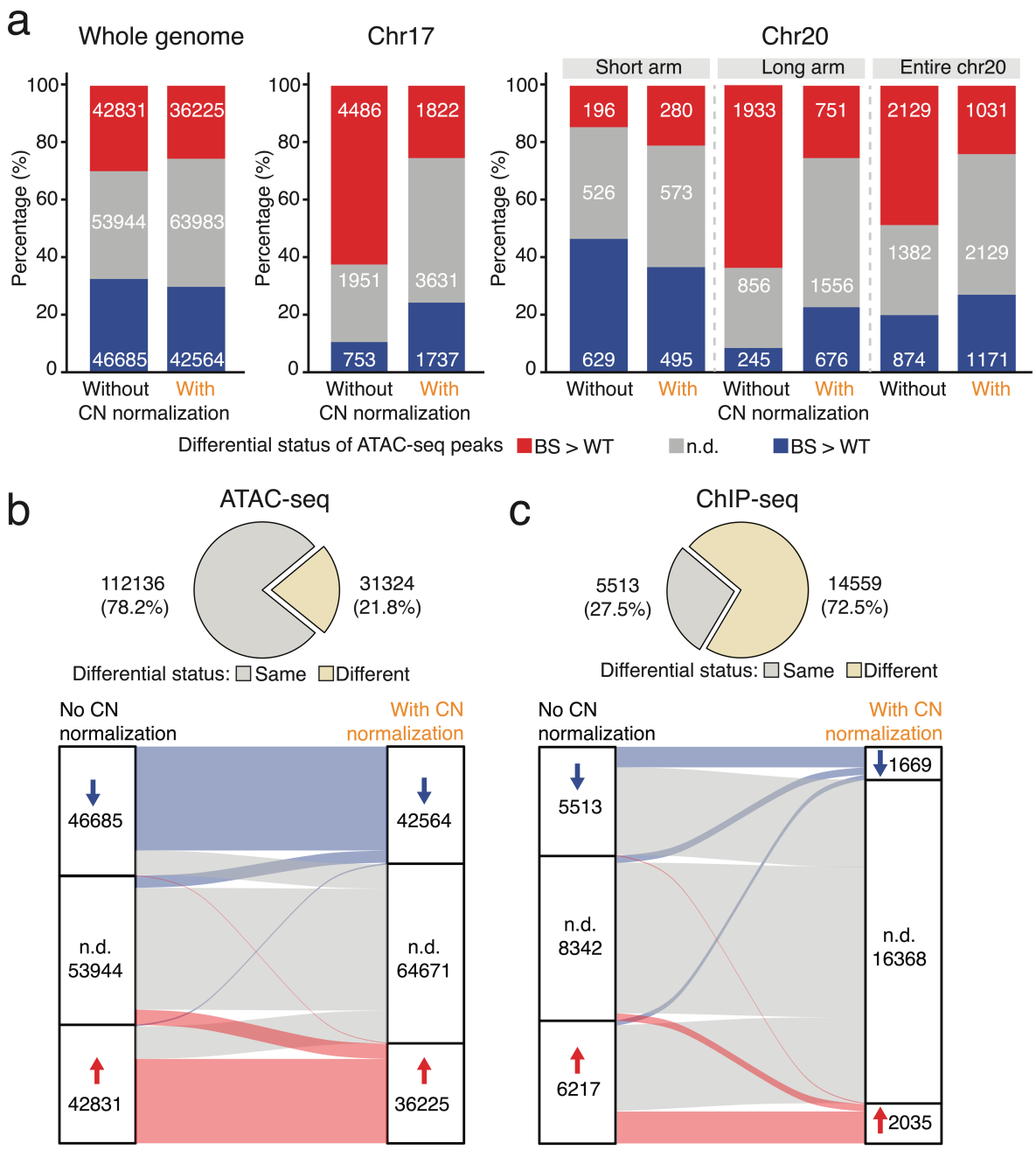



**Figure 3. Impacts of copy number normalization on differential analysis**. (**a**) Number of significantly and not significantly differential ATAC-seq peaks across the genome (left), on chromosome 17 (chr17, middle) and on chr20 (right) before and after applying copy number (CN) normalization. N.d. indicates not significantly differential signals with adjusted *P* (p-adj) ≥ 0.05. (**b**) Differential status of open chromatin regions before and after applying CN normalization. N.d. indicates not significantly differential signals with p-adj ≥ 0.05. (**c**) Differential status of G-quadruplex forming sites before and after applying CN normalization in ChIP-seq data. N.d. indicates not significantly differential signals with FDR ≥ 0.05.

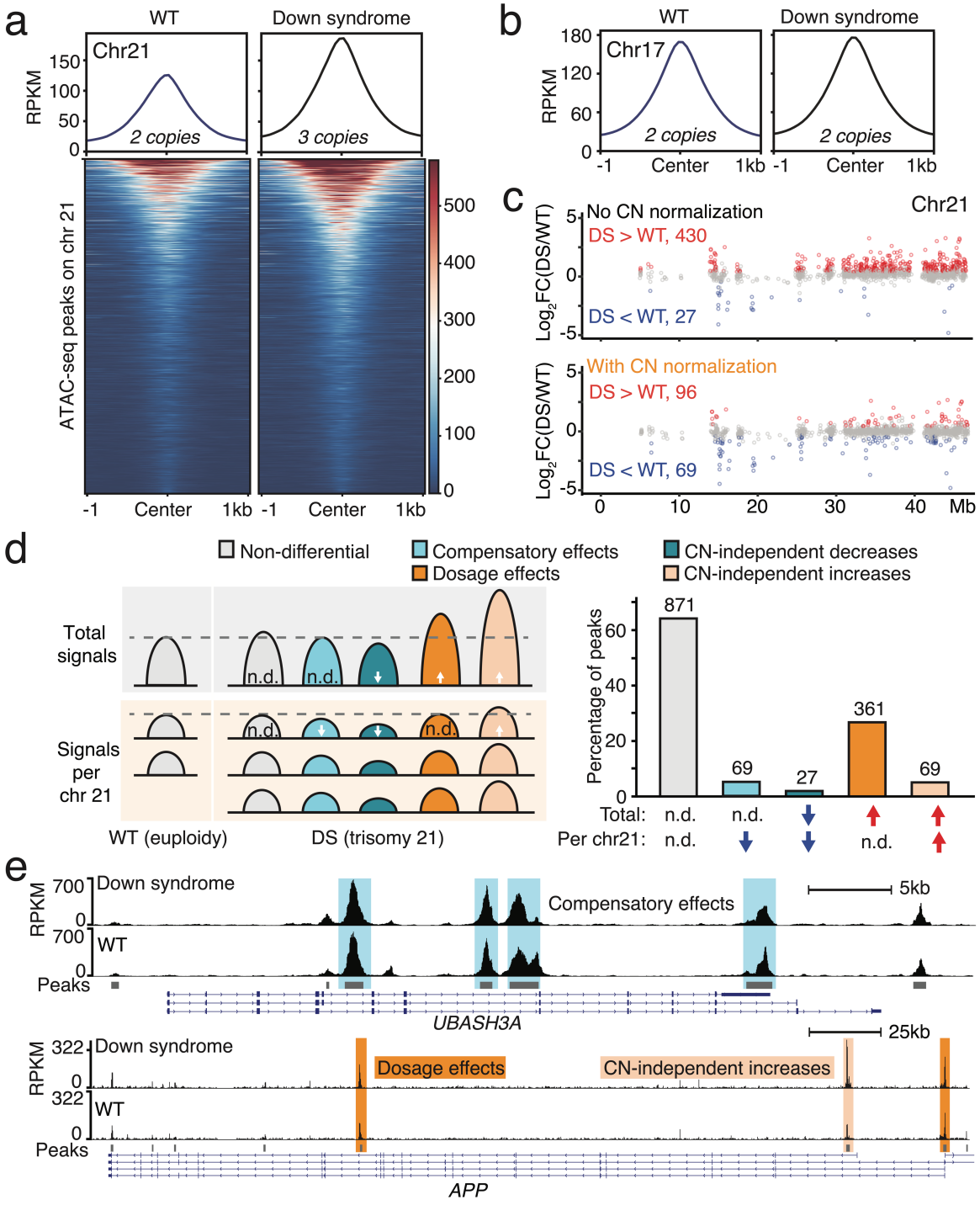
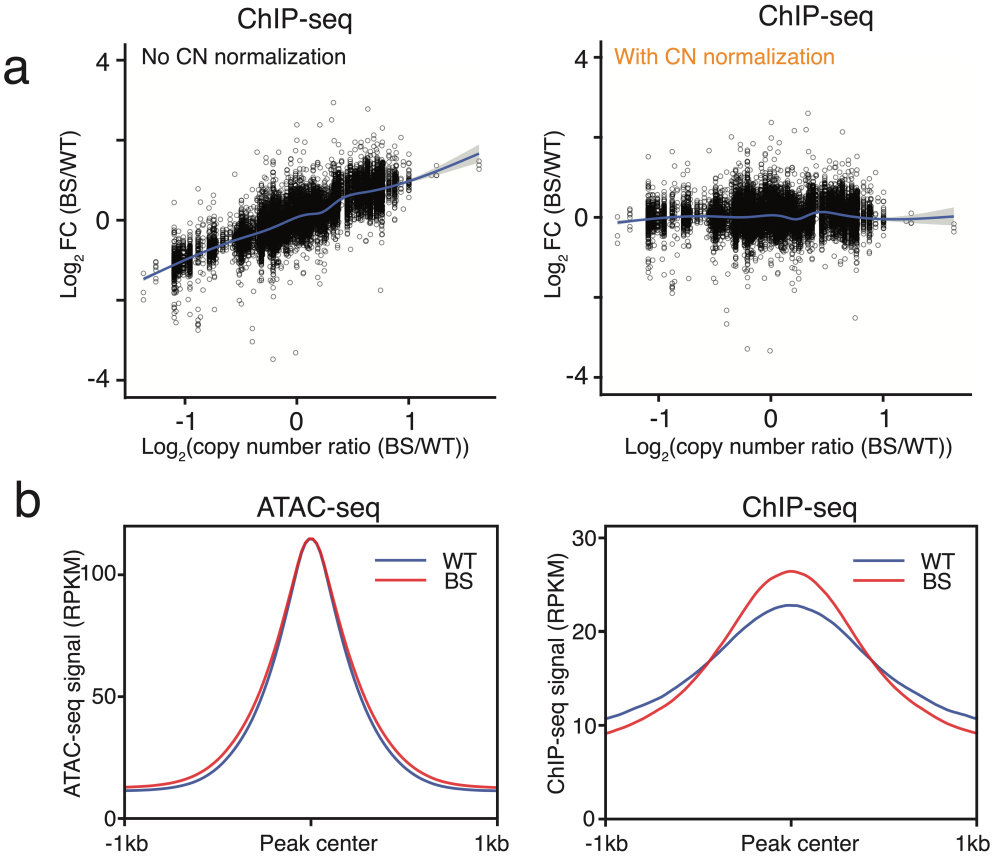
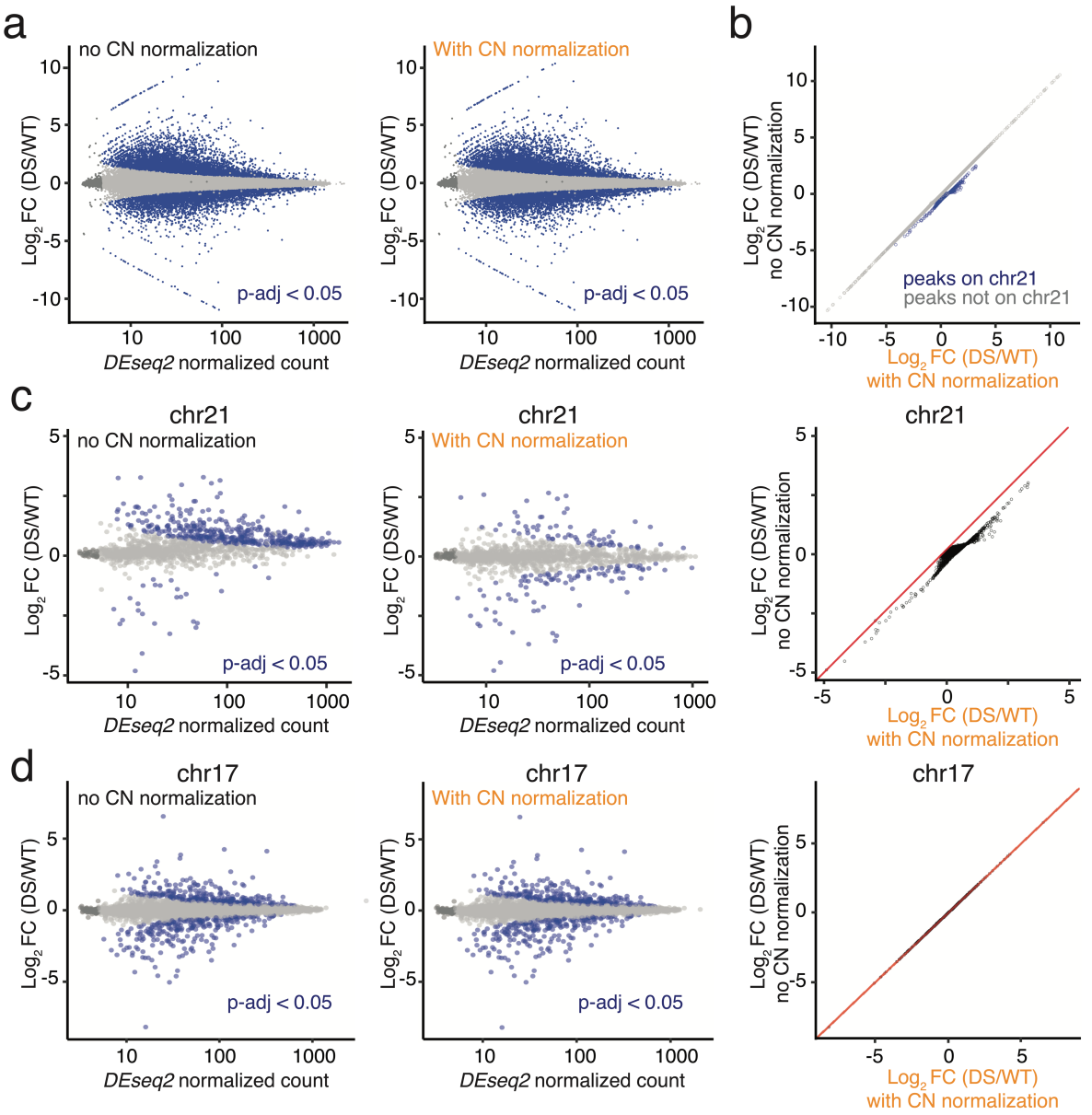



**Figure 4. Copy number normalization identifies regions with dosage and compensatory effects in Down syndrome.** (**a**) Average ATAC-seq signal profiles (top) and chromatin accessibility (bottom) of peaks on chromosome 21 (chr21) in wildtype (WT) and Down syndrome (DS) samples. (**b**) Average ATAC-seq signal profiles of peaks on chr17 in WT and DS samples. (**c**) Distribution of differential signals regarding total chromatin accessibility (top) and average chromatin accessibility per chr21 (bottom) in ATAC-seq peaks. (**d**) Categories of open chromatin regions on chr21 defined according to changes in total chromatin accessibility and average chromatin

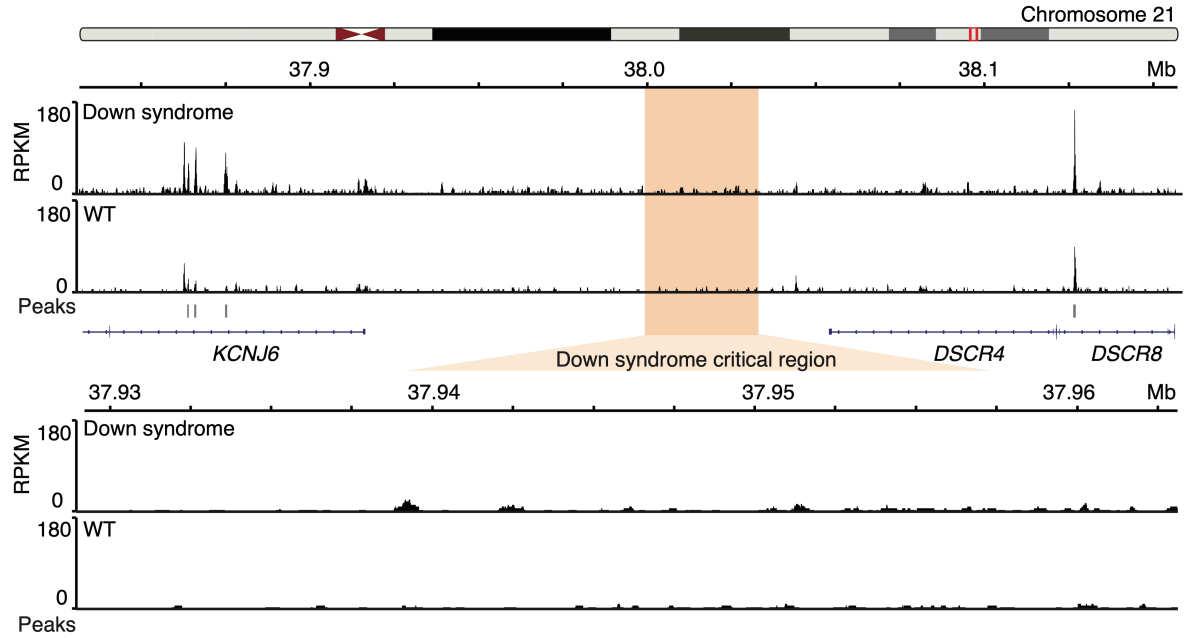
accessibility per chr21. N.d. indicates not significantly differential signals with adjusted *P* (p-adj) < 0.05; arrows denote significantly increased or decreased signals with p-adj < 0.05. (**e**) Open chromatin regions in *UBASH3A* and *APP* genes exhibiting compensatory effects (shaded in blue), dosage effects (shaded in dark orange) or both dosage effects and CN-independent increases (shaded in light orange). Peaks without shading displayed no change in either total or average chromatin accessibility per chr21.



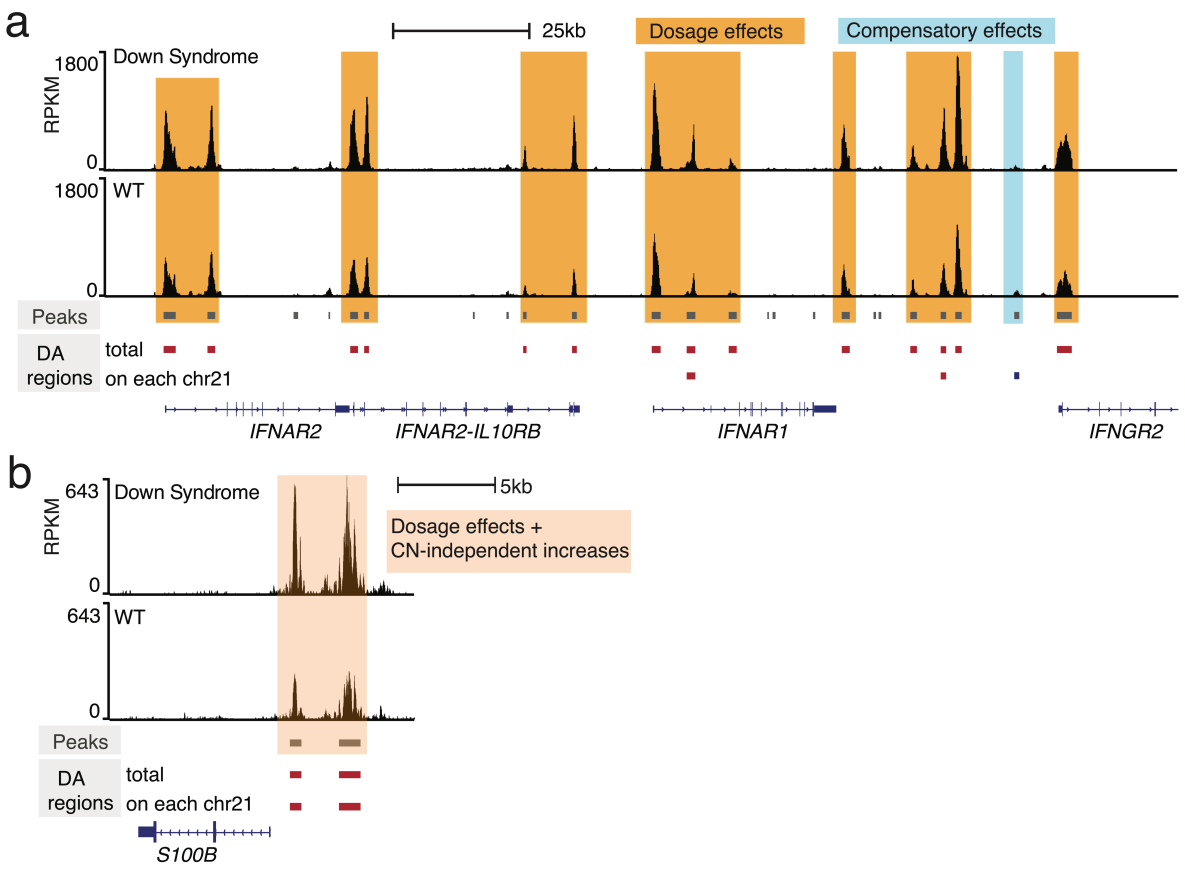

**Supplementary Figure 1. Impacts of copy number normalization on the differential analysis of ATAC-seq and ChIP-seq.** (**a**) The trended biases of differential G-quadruplex formation from copy number differ- ences without (left) and with (right) applying copy number normalization in ChIP-seq data. $Log_2CNR > 0$ and $log_2CNR < 0$ indicate relative number gain and loss in BS, respectively. The lines represent the LOESS smoothing curve for the data with the grey area indicating the 95% confidence interval for the fitted curve. (**b**) Average signal profiles of ATAC-seq and ChIP-seq in their corresponding peak sets.




**Supplementary Figure 2. Impacts of copy number normalization on differential analysis of Down syndrome ATAC-seq data.** (**a**) MA plots of the genome-wide differential signals without (left) and with (right) applying copy number (CN) normalization. Differential peaks with adjusted *P* (p-adj) < 0.05 are depicted in blue while those with p-adj ≥ 0.05 are shown in grey. (**b**) A comparison of the differential signals without and with applying CN normalization for peaks on chromosome 21 (chr21) (blue) and other chromosomes (grey). (**c**) Differential signals for peaks on chr21 without (left) and with (right) CN normalization. (**d**) Differential signals for peaks on chromosome 17 (chr17) without (left) and with (right) CN normalization.




**Supplementary Figure 3. No open chromatin region in the minimal critical region in Down syndrome.** The top panel is genome browser screenshot of ATAC-seq tracks in genes near the Down syndrome critical region. Bottom panel is the zoom-in view of this regions without any ATAC-seq peak.




**Supplementary Figure 4. Example loci on chromosome 21 with dosage and compensatory effects in Down syndrome.** (**a**) The genome browser screenshot of ATAC-seq tracks in interferon receptor gene clusters. Most of them showed increased total chromatin accessibility (shaded in dark orange) and one showed compensatory effects (shaded in blue). Differentially accessible (DA) regions depicted in red and blue bars represent regions with significantly increased and decreased chromatin accessibility, respectively. (**b**) The genome browser screenshot of ATAC-seq tracks near the transcription start site of *S100B* gene. Both showed increased chromatin accessibility in total signal and averaged signal on each chromosome 21. Red bars indicate regions with significantly increased chromatin accessibility in Down syndrome.

# Methods

### Cell culture

Sex-matched and roughly age-matched fibroblast cell lines from Bloom syndrome and healthy donors were obtained from Coriell Institute (BS, GM08505; WT, GM00637). Fibroblast lines were cultured in 1× DMEM (Thermo Scientific, Gibco, Cat. #11960) supplemented with 1% minimum essential medium non-essential amino acids (MEM NEAA; Thermo Scientific, Gibco, Cat. #11140) and 1% Penicillin-Streptomycin (Thermo Scientific, Gibco, Cat. #15140122) and 10% or 15% fetal bovine serum (FBS; Thermo Scientific, Gibco, Cat. #10500064), respectively. Cell passaging was performed according to recommendations by Coriell Institute. All cells were cultured at 37°C under 5% $CO_2$.

To collect cells, fibroblasts were first trypsinized (Trypsin-EDTA solution, Sigma-Aldrich, Cat. #T4049), pelleted at 300 g for 5 min and washed once with cold 1× Dulbecco's phosphate-buffered saline (PBS, Thermo Scientific, Gibco, Cat. #14190).

### ATAC-seq library preparation

Cells were collected as described above and counted (automated cell counter Countess 3, Thermo Scientific; Countess™ Cell Counting Chamber Slides, Thermo Scientific, Cat. #C10283). For each reaction, 50,000 cells were subjected to omni-ATAC-seq preparation as described previously with a modified tagmentation reaction [62]. Nuclei were resuspended in 50 µl of tagmentation mix consisting of 10 µl 5× TAPS-DMF buffer (50 mM TAPS, 25 mM $MgCl_2$, 50% (v/v) DMF), 3 µl Tn5 transposase [63], 16.5 µl 1x PBS, 0.25 µl 2% Digitonin, 0.5 µl 10% (v/v) Tween 20 and 19.75 µl $H_2O$. DNA purified from the transposed product was then amplified via PCR (10 cycles) using Q5 High-Fidelity DNA Polymerase (New England Biolabs, Cat. #M0491L).

### G-quadruplex ChIP-seq protocol and library preparation

***Chromatin preparation.*** Chromatin preparation was performed essentially as previously described with a few modifications [29]. Cells from two full 15cm dishes were initially rinsed once in 1× PBS and subsequently fixed in the dish with 1% formaldehyde in cell culture medium at room temperature for 8.5 minutes. Nuclei were then lysed in 25 µl lysis buffer and sheared in small tubes (microTUBE AFA Fiber Pre-Slit Snap-cap 6×16mm, Covaris, Cat. #520045) using an S220 instrument (Covaris). The shearing parameters were as follows: *water level = 15*, *duty cycle = 15*, *intensity = 6*, and *cycle per burst = 200*. Each chromatin sample underwent shearing for a total of 3 minutes, with a 30-second pause for every one minute of shearing.

***G-quadruplex ChIP-seq.*** For each sample, 2 biological replicates, each of which with 3 technical replicates of immune-precipitation (IP) were prepared. The IP reaction was performed as described previously [29], with the following modifications: during IP, chromatin was incubated with the antibody called BG4 (Absolute antibody, Cat. #Ab00174-30.146) against DNA G-quadruplexes structures for an extended duration of 2 hours, and 10ul of Anti-FLAG M2 Magnetic Beads (Sigma-Aldrich, Cat. #M8823) was used for each IP reaction to pull

down fragmented chromatin with G-quadruplex structures. To purify DNA from both the IP samples (DNA pulled down by the beads) and the input samples (sheared chromatin without going through IP steps) after 5 times of washing the beads with WASH buffer (100 mM KCl, 0.1% (v/v) Tween 20, 10 mM Tris, pH 7.4) to remove residual nonspecific bound chromatin, 75 µl of reverse-crosslinking buffer (0.2% SDS, 1× TE, and 50 mM NaCl) was introduced. The samples were incubated at 37°C for one hour, followed by an overnight incubation at 65°C. After an additional hour of proteinase K digestion at 65°C, DNA purification was carried out using a MinElute kit (QIAGEN, Cat. #28006). The subsequent steps for library preparation were carried out with the DNA Thruplex kit (Takara, Cat. #R400674, Cat. #R400665) following the manufacturer's protocols.

***G-quadruplex ChIP-qPCR and library preparation.*** Prior to library preparation, qPCR was performed to evaluate G-quadruplex enrichment in IP vs. input (DNA extracted from sonicated chromatin without IP). We used 2X CFX SYBR Mix (Applied Biosystems, Cat. #4472942) and a set of primers targeting known G4 positive and negative control sites [29].

**Sequencing and data analysis**

One biological replicate of G4 ChIP-seq libraries was sequenced on a HiSeq3000 platform (150 bp paired-end sequencing; Illumina Inc.; the Genome Core Facility at the Max Planck Institute for Biology, Tübingen). ATAC-seq libraries and the second biological replicate of G4 ChIP-seq libraries were sequenced on a NovaSeq6000 platform (150 bp paired-end sequencing; Illumina Inc.; service provider: GENEWIZ, Leipzig).

A more detailed step-by-step description and pipeline can be found both in the additional file 3 and on GitHub, https://github.com/Dingersrun/Copy-number-normalization

***Read alignment and read filtering.*** Raw FASTQ reads were extracted with *bcl2fastq* (version 2.20). Tn5 and TruSeq adapter sequences, G repeats due to two-color base calling errors and reads shorter than 20 bp were trimmed and removed using *cutadapt* (version 4.0) [64]. Reads were then aligned to the human reference genome version hg38 using *bwa mem* (version: 0.7.17-r1188). Duplicates were marked and removed using *Picard* (https://broadinstitute.github.io/picard/, version 2.18.25). Prior to calling peaks, reads mapped to mitochondrial, reads with a mapping quality lower than 20 and reads in hg38 blacklisted regions (downloaded from http://mitra.stanford.edu/kundaje/akundaje/release/blacklists/hg38human/hg38.blacklist.bed.gz) were discarded.

***Peak calling and differential analysis for ATAC-seq.*** ATAC-seq peak calling was performed using *MACS2* (version 2.1.1.20160309) with *callpeak –format BAMPE --nomodel –min-length 100 narrowPeak* parameters using a bamfile made from pooling equal number reads from each sample [10,65]. The number of fragments in each ATAC-seq peak was counted with *htseq-count* (*HTSeq* version 0.9.1) [15]. Subsequently, the count matrix was normalized and differential analysis was carried out in *DESeq2* (version 1.30.1) [31]. In the case of applying copy number normalization, the count matrix was adjusted before data normalization as described in the following section.

***Copy number normalization.*** A detailed step-by-step description can be found in the additional file 3.

The first step of copy number normalization is characterizing CNR. Genomic sequencing data from cell lines GM08505 (BS) and GM00637 (WT) were used in *CNVkit* following its recommended copy number calling pipeline *(--method wgs --target-avg-size 50000*) to identify copy number alteration in BS relative to WT [30]. Gapped regions in the human genome assembly marked by Ns (http://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/gap.txt.gz) were excluded from the analysis. In the output from *CNVkit*, $\log_2$-transformed values representing CNR for genomic segments were retrieved and converted back to the original value. The second step is assigning each peak to its overlapping DNA segment or the closest DNA segment. The CNR of this segment was then used as a scaling factor to modify the read/fragment count in this peak. Specifically, for the BS vs. WT comparison, if the CNR is greater than or equal to 1 (CNR ≥ 1), the fragment counts in peaks in BS were divided by the CNR. Conversely, if the CNR was smaller than 1 (CNR < 1), the fragment counts in peaks in WT were multiplied by the CNR. For DS vs. WT comparison, the fragment counts in peaks for peaks on chromosome 21 were simply divided by 1.5 in the DS sample.

***Peak calling and differential analysis for ChIP-seq*.** For each biological replicate, a pooled file was generated by subsampling 35 million reads from each technical replicate. The pooled file was subjected to peak calling using *MACS2* (version 2.1.1.20160309) with *callpeak –format BAM narrowPeak* [65]. Peak sets from 2 biological replicates were ranked by their signal values and then filtered based on reproducibility using *IDR* (version 2.0.3) [66]. The final high confidence peaks for each sample were obtained with the criteria of irreproducible discovery rate < 0.05. Peaks were also called with input samples using the same parameters, which were later used as regions to be excluded in the differential analysis in *DiffBind* (version 3.0.15) [18,19]. The differential analysis was carried out by using default parameters with *summits=FALSE, bFullLibSize=FALSE* and *edgeR* normalization [67]. In the case of applying copy number normalization, the read counts in peaks for each sample were extracted from *DiffBind* and modified in the same way as it was for ATAC-seq. A more detailed step-by-step description can be found in supplementary files.

## Supplementary Information

Additional file 1. Copy number ratio in Bloom syndrome cell line relative to wild type cell line for each chromosome. Available at: https://doi.org/10.1101/2024.04.11.58881

Additional file 2. Supplementary figures.

Additional file 3. Step-by-step pipelines for differential analysis with and without copy number normalization for ATAC-seq and ChIP-seq data. Available at: https://doi.org/10.1101/2024.04.11.58881

## Acknowledgments

We thank Detlef Weigel and Marja Timmermans for inspiring this idea. We thank members of the Chan and Jones lab and Yinan Wang for helpful discussions and critical reading of the manuscript. We also thank the Genome Center at the Max Planck Institute for Biology Tübingen for providing support. We thank Andre Noll for computing support.

## Funding

D.S. is supported by an International Max Planck Research School fellowship. The research was supported by the Max Planck Society.

## Availability of data and materials

ATAC-seq and ChIP-seq data from WT and BS fibroblast cell lines are available at the NCBI GEO repository under accession numbers GSE259257 and GSE259258.

ATAC-seq data from Down syndrome and WT are ATAC-seq data obtained from samples cultivated at 37 °C in the study from Cardiello *et al.* under GEO accession number GSE173536 with the following entries: GSM5269423, GSM5269425, GSM5269427, GSM5269428, GSM5269429, GSM5269430, GSM5269433 and GSM5269434 [39].

## Declarations

### Ethics approval

Not applicable.

### Competing interest

The authors declare no competing interest.

## Author details

1, Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2, University of Groningen, Groningen Institute of Evolutionary Life Sciences (GELIFES), 9747 AG Groningen, The Netherlands
* Correspondence to dingwen.su@tuebingen.mpg.de; frank.chan@rug.nl

## References

1. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213–1218 (2013).
2. Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nat Rev Genet* **20**, 207–220 (2019).
3. Ma, S. & Zhang, Y. Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Mol Biomed* **1**, 9 (2020).
4. Furey, T. S. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nat Rev Genet* **13**, 840–852 (2012).
5. Park, P. J. ChIP–seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669–680 (2009).
6. Corces, M. R. *et al.* The chromatin accessibility landscape of primary human cancers. *Science* **362**, eaav1898 (2018).
7. Wang, J. *et al.* ATAC-Seq analysis reveals a widespread decrease of chromatin accessibility in age-related macular degeneration. *Nat Commun* **9**, 1364 (2018).
8. Yan, F., Powell, D. R., Curtis, D. J. & Wong, N. C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biol* **21**, 22 (2020).
9. Reske, J. J., Wilson, M. R. & Chandler, R. L. ATAC-seq normalization method can significantly affect differential accessibility analysis and interpretation. *Epigenetics & Chromatin* **13**, 22 (2020).
10. Gaspar, J. M. *Improved Peak-Calling with MACS2*. http://biorxiv.org/lookup/doi/10.1101/496521 (2018) doi:10.1101/496521.
11. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res* **22**, 1813–1831 (2012).
12. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines. 2023.04.04.535623 Preprint at https://doi.org/10.1101/2023.04.04.535623 (2023).
13. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
14. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Research* **42**, W187–W191 (2014).
15. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
16. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
17. Lun, A. T. L. & Smyth, G. K. csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Research* **44**, e45–e45 (2016).
18. Stark, R. & Brown, G. DiffBind: Differential binding analysis of ChIP-Seq peak data.
19. Ross-Innes, C. S. *et al.* Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* **481**, 389–393 (2012).

20. Liang, Q., Conte, N., Skarnes, W. C. & Bradley, A. Extensive genomic copy number variation in embryonic stem cells. *Proceedings of the National Academy of Sciences* **105**, 17453–17456 (2008).
21. Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984–991 (2022).
22. Shadeo, A. & Lam, W. L. Comprehensive copy number profiles of breast cancer cell model genomes. *Breast Cancer Research* **8**, R9 (2006).
23. Stepanenko, A. A. & Dmitrenko, V. V. HEK293 in cell biology and cancer research: phenotype, karyotype, tumorigenicity, and stress-induced genome-phenotype evolution. *Gene* **569**, 182–190 (2015).
24. Chu, W. K. & Hickson, I. D. RecQ helicases: multifunctional genome caretakers. *Nat Rev Cancer* **9**, 644–654 (2009).
25. Croteau, D. L., Popuri, V., Opresko, P. L. & Bohr, V. A. Human RecQ Helicases in DNA Repair, Recombination, and Replication. *Annu. Rev. Biochem.* **83**, 519–552 (2014).
26. Cunniff, C., Bassetti, J. A. & Ellis, N. A. Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol Syndromol* **8**, 4–23 (2017).
27. Chaganti, R. S. K., Schonberg, S. & German, J. A Manyfold Increase in Sister Chromatid Exchanges in Bloom's Syndrome Lymphocytes. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4508–4512 (1974).
28. Chester, N., Babbe, H., Pinkas, J., Manning, C. & Leder, P. Mutation of the Murine Bloom's Syndrome Gene Produces Global Genome Destabilization. *Molecular and Cellular Biology* **26**, 6713–6726 (2006).
29. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* **13**, 551–564 (2018).
30. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873 (2016).
31. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
32. Hänsel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* **18**, 279–284 (2017).
33. Sun, H., Karow, J. K., Hickson, I. D. & Maizels, N. The Bloom's Syndrome Helicase Unwinds G4 DNA. *Journal of Biological Chemistry* **273**, 27587–27592 (1998).
34. Chester, N., Kuo, F., Kozak, C., O'Hara, C. D. & Leder, P. Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. *Genes Dev.* **12**, 3382–3393 (1998).
35. van Wietmarschen, N. *et al.* BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat Commun* **9**, 271 (2018).
36. Shabtai, F. & Halbrecht, I. Bloom's syndrome, missing Y, hypogonadism and cancer. *Clinical Genetics* **18**, 93–95 (1980).
37. Antonarakis, S. E. *et al.* Down syndrome. *Nat Rev Dis Primers* **6**, 9 (2020).
38. Wilcock, D. M. & Griffin, W. S. T. Down's syndrome, neuroinflammation, and Alzheimer neuropathogenesis. *J Neuroinflammation* **10**, 864 (2013).

39. Cardiello, J. F., Westfall, J., Dowell, R. & Allen, M. A. *Characterizing Primary Transcriptional Responses to Short Term Heat Shock in Paired Fraternal Lymphoblastoid Lines with and without Down Syndrome*. http://biorxiv.org/lookup/doi/10.1101/2023.01.17.524431 (2023) doi:10.1101/2023.01.17.524431.
40. Pelleri, M. C. *et al.* Systematic reanalysis of partial trisomy 21 cases with or without Down syndrome suggests a small region on 21q22.13 as critical to the phenotype. *Hum. Mol. Genet.* ddw116 (2016) doi:10.1093/hmg/ddw116.
41. Hwang, S. *et al.* Consequences of aneuploidy in human fibroblasts with trisomy 21. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2014723118 (2021).
42. Waugh, K. A. *et al.* Triplication of the interferon receptor locus contributes to hallmarks of Down syndrome in a mouse model. *Nat Genet* **55**, 1034–1047 (2023).
43. Liu, S. *et al.* Aneuploidy effects on human gene expression across three cell types. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218478120 (2023).
44. Stamoulis, G. *et al.* Single cell transcriptome in aneuploidies reveals mechanisms of gene dosage imbalance. *Nat Commun* **10**, 4495 (2019).
45. Antonarakis, S. E. Down syndrome and the complexity of genome dosage imbalance. *Nat Rev Genet* **18**, 147–163 (2017).
46. M, S. *et al.* Impact of increased APP gene dose in Down syndrome and the Dp16 mouse model. *Alzheimer's & dementia : the journal of the Alzheimer's Association* **18**, (2022).
47. Conrad, T. & Akhtar, A. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**, 123–134 (2012).
48. Xing, Z. *et al.* Dissection of a Down syndrome-associated trisomy to separate the gene dosage-dependent and -independent effects of an extra chromosome. *Human Molecular Genetics* **32**, 2205–2218 (2023).
49. Chung, H., Green, P. H. R., Wang, T. C. & Kong, X.-F. Interferon-Driven Immune Dysregulation in Down Syndrome: A Review of the Evidence. *J Inflamm Res* **14**, 5187–5200 (2021).
50. Malle, L. *et al.* Autoimmunity in Down's syndrome via cytokines, CD4 T cells and CD11c+ B cells. *Nature* **615**, 305–314 (2023).
51. Illouz, T. *et al.* Immune Dysregulation and the Increased Risk of Complications and Mortality Following Respiratory Tract Infections in Adults With Down Syndrome. *Front. Immunol.* **12**, 621440 (2021).
52. Sullivan, K. D. *et al.* Trisomy 21 consistently activates the interferon response. *eLife* **5**, e16220 (2016).
53. Malle, L. *et al.* Excessive negative regulation of type I interferon disrupts viral control in individuals with Down syndrome. *Immunity* **55**, 2074-2084.e5 (2022).
54. Ge, Y., Paisie, T. K., Chen, S. & Concannon, P. UBASH3A Regulates the Synthesis and Dynamics of TCR–CD3 Complexes. *The Journal of Immunology* **203**, 2827–2836 (2019).
55. Kuilman, T. *et al.* CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol* **16**, 49 (2015).
56. Scheinin, I. *et al.* DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. *Genome Res.* **24**, 2022–2032 (2014).
57. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

58. Robinson, M. D. *et al.* Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res.* **22**, 2489–2496 (2012).
59. Qiu, X. *et al.* CoBRA: Containerized Bioinformatics Workflow for Reproducible ChIP/ATAC-seq Analysis. *Genomics, Proteomics & Bioinformatics* **19**, 652–661 (2021).
60. Ashoor, H. *et al.* HMCan: a method for detecting chromatin modifications in cancer samples using ChIP-seq data. *Bioinformatics* **29**, 2979–2986 (2013).
61. Ashoor, H., Louis-Brennetot, C., Janoueix-Lerosey, I., Bajic, V. B. & Boeva, V. HMCan-diff: a method to detect changes in histone modifications in cells with different genetic characteristics. *Nucleic Acids Res* gkw1319 (2017) doi:10.1093/nar/gkw1319.
62. Corces, M. R. *et al.* An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. 15 (2017).
63. Picelli, S. & Sandberg, R. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. 9.
64. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j.* **17**, 10 (2011).
65. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* **9**, R137 (2008).
66. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, (2011).
67. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR : a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

# Discussion

Previous studies on the association of G4 and differentially expressed genes and SCEs in BS were limited to *in silico* predicted G4 motifs or *in vitro* validated G4-seq hits. Crucially, the presence of G4-forming sequences poorly predicts the G4 formation *in vivo*. This thesis is the first study to map endogenous G4 in BS cells, providing direct evidence and highlighting the central role of G-quadruplexes, a previously underappreciated factor and the substrate of BLM helicase, in the observed molecular changes in BS cells and BS individuals. I demonstrated a significant association and partial causation between local changes in G4 formation and epigenetic and transcriptomic changes in BS cells. Furthermore, I dissected the plausible molecular mechanism of G4 regulating downstream molecular changes in BS and proposed a molecular model of G4 in the etiology of BS.

**The association between the local presence of G4 and SCEs in BS**

The elevated frequency of SCEs serves as the signature and diagnostic criterion of BS. Strand-Seq, a single-cell template strand sequencing method, allows mapping SCE events in individual cells at approximately 10kb resolution [144]. With this method, Van Wietmarschen et al. characterized SCEs at approximately 10 kb resolution in eight cell lines, four derived from healthy donors and four from BS individuals. They observed SCEs occurring at common fragile site hotspots in both WT and BS cells, albeit with no stronger enrichment from BS cells. However, leveraging *in silico* predicted G4-motifs, they noted a significant enrichment of SCEs overlapping with G4 motifs in BS, but not in WT cells [25]. Notably, one of the BS cell lines used in their study was also employed in my experiments.

Utilizing G4 sites mapped via G4 ChIP-seq in the same BS cell line, I demonstrated significant enrichment of SCEs at endogenous G4 sites, with approximately 20% more overlap with G4 peaks than expected by chance. Additionally, by intersecting SCEs with ATAC-seq peaks, I found they were enriched at open chromatin regions as well. Van Wietmarschen et al. also noted the significant enrichment of SCE regions in actively transcribed genes in all BS cell lines, but not in WT cells [25]. I directly showed

that in general G4 peaks are also enriched in active genes and mostly reside in open chromatin regions. Together, our results strongly support the hypothesis that SCEs in BS may be induced via G4, providing robust evidence for the model that, in the absence of BLM, G4 structures are not effectively resolved and are thus more likely to persist, thereby impeding DNA replication machinery progression and stalling DNA polymerase. To restore the replication fork, the HR DNA repair pathway utilizes the sister chromatid as the template, leading to the formation of dHJs. Importantly, due to the absence of BLM, dHJs cannot be dissolved by the BTR complex, resulting in increased SCE events (**Fig. 5** in Introduction).

**Enrichment of endogenous G4 in differentially expressed genes in BS**

Several studies have reported the association of G4-forming sequences with differentially expressed genes in BS. Specifically, G4 motifs were found to be enriched at differentially expressed genes in BLM-deficient fibroblast cell lines, particularly at the junction site of the first exon and the first intron on the non-transcribing strand [18,26]. However, another study expanded the sample size and found that the observed enrichment is attributed to the GC content in these regions. Upon GC content correction, there is no significant enrichment of G4-forming sequences using G4-seq hits [27]. In this study, I investigated this association with G4 peaks and observed that G4 peaks were generally enriched in gene bodies, particularly in the 5'UTR, promoter, the first exon, and the first intron, as well as the 1st-Ext-Int junction sites. Compared to the background enrichment, DE genes in the two BS cell lines, G4 peaks did not show a consistent enrichment at the transcription start site (TSS) or at the previously claimed 1st-Ext-Int junction sites. Notably, there was a significant correlation between changes in G4 formation and changes in gene expression, suggesting that it is not the local presence of G4s, but rather changes in their activities, that are more relevant to BS.

**Reciprocal modulation of G4 formation and chromatin accessibility**

I found that G4 ChIP-seq and ATAC-seq exhibited highly similar profiles in promoter regions. Across all samples, G4 peaks predominantly resided in open chromatin regions and in regions with denser ATAC-seq peaks. Notably, the presence of G4 was

associated with wider and more accessible open chromatin regions. Moreover, in BS cells of both cell types, there was also a strong positive correlation between the directional changes of G4 formation and chromatin accessibility: regions with increased G4 formation activity mostly exhibited increased chromatin accessibility, and vice versa. The observed coherence and strong correlation in the changes between ATAC-seq and G4 ChIP-seq were mainly confined to the ATAC-seq peaks overlapping G4 peaks, and strongly distorted between G4 peaks and proximal ATAC-seq peaks. These findings suggest that changes in G4 formation could feedback and modulate focal chromatin accessibility, which echoes the findings from other studies. Prorok et al. demonstrated that upon genetically abolishing the G4 motif in the *MYC* gene by CRISPR genome editing, there is a novel nucleosome positioning at the mutant G4 site [144]. Similarly, Li et al. demonstrated that in some genes treatment of a G4 stabilizing molecule increased the chromatin accessibility in their promoters [139]. Although the open chromatin state provides a favorable environment for G4 to form, the formation of G4 possibly makes the regions more accessible by displacing histones, which leads to stronger and wider signals in ATAC-seq and effectively creates "space" for other factors to come in.

**Impact of G4 strandedness on gene expression**

As a broad trend and also in line with previous studies [22,23], genes with a G4 peak in the promoter showed overall higher expression. Further, by intersecting G4 peaks with G4-seq hits, I inferred whether G4s form on the Watson or Crick strand and their strandedness relative to the gene's transcription direction. Approximately half of G4 peaks can be assigned to the transcribing or non-transcribing strand. Interestingly, genes with G4s regardless of the strandedness displayed significantly higher gene expression levels. Particularly, in my data genes on the transcribing strand showed significantly lower expression levels than genes with G4s on the non-transcribing strand. *In vitro* studies support that the strandedness of G4 structures in the promoter matters in regulating gene expression. However, it was found that G4 formation on the non-transcribing strand enhances gene expression, and conversely, G4 formation on the transcribing strand can completely block the transcription if G4 forms at 100% chance and persists [141]. This discrepancy with my observations could be explained by

the fact that G4 ChIP-seq is a bulk assay and it detects the averaged signals across all cells. Having a G4 peak usually represents the G4-forming status in a substantial proportion of cells used for the sample preparation. In the case of a G4 peak in the non-transcribing strand, in this subset of cells, the transcription is suppressed. However, as G4s mostly form in highly expressed genes, the averaged expression level of such genes is lower than genes with G4 on the transcribing strand though, it may still be higher than those genes without G4 peaks.

**Possible molecular mechanisms of G4 regulating gene expression**

Both the changes in ATAC-seq and G4 ChIP-seq were positively correlated with gene expression changes, with their impacts being additive. By focusing on the gene set where signals from ATAC-seq, RNA-seq, and ChIP-seq were all detected, I investigated the potential molecular hierarchy between changes in chromatin accessibility and G4 formation regarding their impact on gene expression. My findings suggest that changes in G4 formation (e.g., increased) likely influence (increase) the local chromatin accessibility, thereby modulating (increasing) gene expression. It has been demonstrated that the formation of G4 in *MYC* promoter alters local nucleosome positioning such that the DNA is accessible to RNA polymerase and transcription factors, which enhances transcription activity [144]. Further studies are needed to gain deeper insight into how G4 features in regulating gene expression.

**G4 stabilization via PDS partially recapitulating molecular changes in BS**

While my results demonstrated that changes in G4 formation were tightly associated with multiple molecular phenotypes observed in BS, G4 stabilization via PDS partially recapitulated the molecular changes in BS. This could be attributed to several reasons. Firstly, PDS treatment is short-term, whereas BS imposes long-term effects on the cell lines. Secondly, unwinding G4 is one of BLM's functions, and G4 stabilization via PDS mimics the universal loss of G4-resolving abilities, not limited to BLM. Other helicases such as WRN, FANCJ, and DHX36 can also unwind DNA G-quadruplexes [151,152]. Moreover, the dynamics of G4 could differ in BS and upon PDS treatment. With PDS, the equilibrium of G4 strongly shifts toward G4 formation, and persistent G4 formation may render the focal chromatin rigid and impose even higher stress on molecular

processes. Despite the aforementioned potential discrepancies between PDS treatment and defective G4-resolving abilities in BS, PDS treatment recapitulated a substantial amount of chromatin accessibility and gene expression changes in BS, suggesting a central role of G4 in BS. For future studies, to make the time-course comparable, PDS treatment can be carried out at lower concentrations and prolonged times; alternatively, ML216, a small molecule that inhibits BLM helicase, can be utilized to characterize the short-term effects of loss-of-function of BLM [153].

**Implementing copy number normalization in the differential analysis for BS vs. WT sample pair**

Although CNV is often overlooked in the differential analyses, I demonstrated that copy number differences between contrasted samples could drive the differential signals in ATAC-seq and ChIP-seq data from the Fib-BS vs. Fib-WT. Specifically, ATAC-seq and ChIP-seq peaks were prone to be identified as increased signals towards the sample with a relatively higher copy number in the corresponding genomic interval.

Importantly, BS is primarily caused by loss-of-function of BLM helicase rather than aneuploidy and in particular, the WT cell lines also showed widespread chromosomal aberrations, therefore I interpret the chromosomal aberrations in the fibroblast cell lines to be an artifact during the immortalization of the cell line rather than a primary effect of BS. With the concept that in ATAC-seq and ChIP-seq, the captured signal in a genomic interval is the aggregated signals from all copies, I developed a pipeline featuring copy number normalization to address this matter in general, as well as to accurately identify the differential signals by the factor of interest, BS. Via copy number normalization, the signals in a genomic interval in the contrasted sample were adjusted to the same copy number before downstream data normalization and detecting differential signals. Applying this, I effectively removed the differential signals in ATAC-seq and ChIP-seq arising from copy number differences, which counts for approximately 20% of differential signals in ATAC-seq and 70% in ChIP-seq data. These false-positive signals would be misattributed to BS, leading to misinterpretation of results.

Of note, a similar trend to the observed impacts of CNV in the ATAC-seq and ChIP-seq data analysis with a copy-number blind pipeline was observed in RNA-seq data. There was an overrepresentation of differentially expressed genes in regions with more pronounced copy number differences. However, due to the highly complex regulatory mechanisms of transcription, I chose not to apply copy number normalization in the RNA-seq data analysis. Applying the concept of copy number normalization in RNA-seq data analysis assumes the linear relationship between copy number and gene expression. However, transcription regulation involves intricate interactions among various factors, including but not limited to copy number, DNA methylation, histone modification, chromatin accessibility, small non-coding RNA, and transcription factors. There were studies showing the increased transcription upon increased copy number in samples with aneuploidy and CNV [148]. Notably, the relationship between copy number and transcription output is rarely linear.

In contrast, the chromatin state seems to be a more independent unit, specifically, given that on the same chromosome the trend of overrepresentation of increased or decreased signals is coherent with the relative local copy number gains or losses. Moreover, in the ATAC-seq and ChIP-seq data, copy number normalization mainly changes the results in regions with copy number differences and very nuanced changes on the differential signals in regions without CNV, suggesting relatively independent and focal impacts of copy number on chromatin signals.

**Identifying open chromatin regions with dosage effects and dosage compensation in Down Syndrome via copy number normalization**

In contrast to the BS, trisomy 21 is the primary cause of Down Syndrome, and thus differential signals driven by the copy number of chromosome 21 are of central interest. In this case, I showed the advantage of combining the conventional copy-number-blind pipeline and my copy-number-aware pipeline. Specifically, for peaks on chromosome 21, the former identified changes in the total signal from all chromosome 21 copies, i.e., CN-dependent changes, and the latter detected the changes in averaged chromatin accessibility on each chromosome 21, i.e., CN-dependent changes. I extended the concept of dosage effects and dosage compensation to chromatin

accessibility. By combining these results for each peak on chromosome 21, peaks with dosage effects can be further divided into two groups, without and with CN-independent increases. Importantly, I revealed the subset of peaks with genuine dosage compensatory effects, whose total chromatin accessibility remained the same whereas the average signal on each chromosome 21 decreased. For instance, I observed chromatin accessibility dosage effects of peaks in many genes known to have transcription dosage effects, e.g., the interferon receptor gene clusters and the risk gene for Alzheimer's disease, *APP* [155]. Interestingly, in *UBASH3A*, a gene related to autoimmune function [156], nearly all the peaks in this gene showed compensatory effects with decreased chromatin accessibility on each chromosome 21, suggesting a critical role and tight regulation of this gene. Such regions with CN-independent changes would have been overlooked without considering CNV, yet they are potentially clinically relevant. Notably, CN-independent changes possibly arise secondarily from the primary molecular changes upon trisomy 21 and can provide novel insights into dosage imbalance and the pathogenesis of DS.

**Opportunities for improvement**

In this study, I used one pair of WT and BS lymphoblastoids and one pair of WT and BS fibroblast cell lines. Although they have been derived from sex-matched and age roughly matched healthy and BS individuals, there are other variables between the samples besides the genotype. For instance, for the Fib pair of samples, I noticed potential biases in the differential analysis driven by copy number variation and corrected it with copy number normalization. Moreover, using samples from two tissue backgrounds alleviates the impacts of sample-specific signals. However, there are other factors like genetic background and the precise tissue for sampling that cannot be fully accounted for in data analysis. To further improve the experimental design, it would be advantageous to expand the sample size, including more pairs of cell lines and even using cell lines whose SCEs have been mapped, use primary fibroblast cell lines without aneuploidy, and generate isogenic BS and WT cell lines via genome editing.

Cleavage under targets and release using nuclease (CUT&RUN) and CUT&Tag were recently developed genomic techniques providing alternatives to the classical ChIP-seq approach [128]. They are simpler, more convenient, require less input material, and importantly, improve signal-to-noise ratios. Meanwhile, in the field of G4, ChIP-seq has gradually been replaced by CUT&Tag using the same antibody, due to its higher sensitivity and lower backgrounds [21,24]. Without the step of fixation, CUT&Tag could capture the G4 in the native state and thus avoid potential biases introduced by cross-linking G4 to chromatin and damage to G4 structures during shearing of the chromatin. With the G4 ChIP-seq, the detected fold changes observed in this study in G4 formation were at small magnitudes, which could be attributed to relatively high background signals. Therefore, mapping endogenous G4 in BS cells with CUT&Tag would benefit the data quality and G4s detected by CUT&Tag could better reflect the G4 profiles in the cells. In addition to G4 ChIP-seq, a straightforward approach to show how BLM deficiency affects the formation of G4 would be performing immunostaining against G4 structures in the cells and quantifying the G4 signals.

To address the causal relationship between G4 and molecular changes in BS, I exclusively utilized one of the G4 ligands, pyridostatin. Exploring the application of other ligands could bolster my hypothesis. However, it's worth noting that the current concentration and duration of PDS treatment had harsh impacts on the cells, particularly on the fibroblasts. Optimizing the concentration through titration with smaller increments and varying treatment durations could have provided further insights.

**Outlook**

In summary, this study has unveiled a novel mechanism of gene regulation involving G4 in BS. The alterations in G4 formation are intricately linked to and partially responsible for the molecular changes observed in BS. By elucidating the regulatory role of G4 structures in the context of BLM deficiency, my work expands our understanding of BLM function beyond its well-established role as a helicase. However, several key questions remain unanswered.

Expanding on this research, a logical next step would involve acquiring the G4 profile after G4 stabilization and assessing its similarity to the G4 profile in BS. Additionally, a rescue experiment would also be necessary in order to fully address the causation. Recently a G4 destabilizing molecule was discovered [157]. Apart from a classical rescue experiment by re-expressing functional BLM in BS cells, this G4 disruptor can be applied to BS cells to see if it can rescue the G4 and other molecular phenotypes. The formation of G4 structures has the potential to influence various aspects of gene regulation. It could change chromatin accessibility, displace factors like TFs, RNA polymerase II, histone modifying enzymes and DNA methyltransferase, and influence three-dimensional chromatin organization, all of which could subsequently modulate gene expression directly or indirectly [144,146,149]. My results suggest that in BS, changes in G4 formation modulate gene expression by regulating chromatin accessibility, but a more detailed molecular mechanism remains unclear. Further studies are required to characterize those related transcription signatures and fill the gap.

BS has complex clinical manifestations across different tissues. Not only did I observe cell-type specific effects of BLM deficiency but also various studies demonstrated that G4 profiles are cell-type specific [22,23,145]. Therefore, generating a comprehensive G4 atlas across different cell types and tissues can help systematically evaluate the role of G4 in the etiology of BS as well as the tissue-specific regulatory changes. Moreover, considering that G4 formation requires G-rich sequences, genetic variation, e.g., single-nucleotide polymorphism (SNPs), among differential individuals may result in intrinsically different G4-forming potentials, which has been shown to be relevant to disease states [158]. Understanding the diverse G4 profiles among tissues and individuals could offer insights into the broad spectrum of clinical symptoms observed in BS individuals and the variability among affected individuals. Future studies in this area will help to unravel the intricate links between genetics, G4 biology, and clinical manifestations, while also assessing the therapeutic potential of targeting G4 structures.

# Glossary

| | |
|---|---|
| °C | Celsius |
| min | Minutes |
| bp | Base pair |
| kb | Kilobase pair |
| Mb | Megabase pair |
| mRNA | Messenger RNA |
| cDNA | Complementary DNA |
| ssDNA | Single-stranded DNA |
| dsDNA | Double-stranded DNA |
| HJ | Holliday Junction |
| G4 | G-quadruplex |
| D-loop | Displacement-loop |
| DSB | DNA double-strand break |
| HR | Homologous recombination |
| NHEJ | Non-homologous end joining |
| MMEJ | Microhomology-mediated end joining |
| SDSA | Synthesis-dependent strand annealing |
| SSA | Single-strand annealing |
| SCE | Sister chromatid exchange |
| LOH | Loss of heterozygosity |
| BS | Bloom Syndrome |
| DS | Down syndrome |
| WT | Wildtype |
| GO | Gene Ontology |
| NGS | Next-generation sequencing |
| DNA | Deoxyribonucleic acid |
| RNA | Ribonucleic acid |
| CNV | Copy number variation |
| CNR | Copy number ratio |
| DA | Differential anlaysis |
| ATAC-seq | Assay for Transposase-Accessible Chromatin using sequencing |
| RNA-seq | RNA sequencing |
| ChIP-seq | Chromatin immunoprecipitation followed by sequencing |
| Strand-seq | Single-cell DNA template strand sequencing |
| CUT&Tag | Cleavage Under Targets and Tagmentation |
| IP | Immunoprecipitation |
| FC | Fold change |
| n.d. | non-differential |
| sig. diff. | Significantly differential ($P < 0.05$) |
| t.s. | transcribing strand |
| TSS | transcription start site |

# Reference


1. Ellis, N. A. *et al.* The Bloom's syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655–666 (1995).
2. Bloom, D. Congenital telangiectatic erythema resembling lupus erythematosus in dwarfs; probably a syndrome entity. *AMA Am J Dis Child* **88**, 754–758 (1954).
3. Hickson, I. D. RecQ helicases: caretakers of the genome. *Nat Rev Cancer* **3**, 169–178 (2003).
4. Chu, W. K. & Hickson, I. D. RecQ helicases: multifunctional genome caretakers. *Nat Rev Cancer* **9**, 644–654 (2009).
5. Croteau, D. L., Popuri, V., Opresko, P. L. & Bohr, V. A. Human RecQ Helicases in DNA Repair, Recombination, and Replication. *Annu. Rev. Biochem.* **83**, 519–552 (2014).
6. Sun, H., Karow, J. K., Hickson, I. D. & Maizels, N. The Bloom's Syndrome Helicase Unwinds G4 DNA. *J. Biol. Chem.* **273**, 27587–27592 (1998).
7. Krejci, L., Altmannova, V., Spirek, M. & Zhao, X. Homologous recombination and its regulation. *Nucleic Acids Research* **40**, 5795–5818 (2012).
8. Bizard, A. H. & Hickson, I. D. The Dissolution of Double Holliday Junctions. *Cold Spring Harbor Perspectives in Biology* **6**, a016477–a016477 (2014).
9. Cunniff, C., Bassetti, J. A. & Ellis, N. A. Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol Syndromol* **8**, 4–23 (2017).
10. Chester, N., Kuo, F., Kozak, C., O'Hara, C. D. & Leder, P. Stage-specific apoptosis, developmental delay, and embryonic lethality in mice homozygous for a targeted disruption in the murine Bloom's syndrome gene. *Genes & Development* **12**, 3382–3393 (1998).
11. Luo, G. *et al.* Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nat Genet* **26**, 424–429 (2000).
12. German, J., Bloom, D. & Passarge, E. Bloom's syndrome. V. Surveillance for cancer in affected families. *Clinical Genetics* **12**, 162–168 (1977).
13. Chaganti, R. S., Schonberg, S. & German, J. A manyfold increase in sister chromatid exchanges in Bloom's syndrome lymphocytes. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4508–4512 (1974).
14. LaRocque, J. R. *et al.* Interhomolog recombination and loss of heterozygosity in wild-type and Bloom syndrome helicase (BLM)-deficient mammalian cells. *Proceedings of the National Academy of Sciences* **108**, 11971–11976 (2011).
15. Nichols, C. A. *et al.* Loss of heterozygosity of essential genes represents a widespread class of potential cancer vulnerabilities. *Nat Commun* **11**, 2517 (2020).
16. Chester, N., Babbe, H., Pinkas, J., Manning, C. & Leder, P. Mutation of the murine Bloom's syndrome gene produces global genome destabilization. *Mol Cell Biol* **26**, 6713–6726 (2006).
17. Dutertre, S. *et al.* Cell cycle regulation of the endogenous wild type Bloom's syndrome DNA helicase. *Oncogene* **19**, 2731–2738 (2000).
18. Nguyen, G. H. *et al.* Regulation of gene expression by the BLM helicase correlates with the presence of G-quadruplex DNA motifs. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9905–9910 (2014).

19. Puig Lombardi, E. & Londoño-Vallejo, A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Research* **48**, 1–15 (2020).
20. Huppert, J. L. & Balasubramanian, S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research* **33**, 2908–2916 (2005).
21. Hänsel-Hertsch, R., Spiegel, J., Marsico, G., Tannahill, D. & Balasubramanian, S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nat Protoc* **13**, 551–564 (2018).
22. Hänsel-Hertsch, R. *et al.* G-quadruplex structures mark human regulatory chromatin. *Nat Genet* **48**, 1267–1272 (2016).
23. Lago, S. *et al.* Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome. *Nat Commun* **12**, 3885 (2021).
24. Lyu, J., Shao, R., Kwong Yung, P. Y. & Elsässer, S. J. Genome-wide mapping of G-quadruplex structures with CUT&Tag. *Nucleic Acids Research* **50**, e13–e13 (2022).
25. van Wietmarschen, N. *et al.* BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat Commun* **9**, 271 (2018).
26. Johnson, J. E., Cao, K., Ryvkin, P., Wang, L.-S. & Johnson, F. B. Altered gene expression in the Werner and Bloom syndromes is associated with sequences having G-quadruplex forming potential. *Nucleic Acids Res* **38**, 1114–1122 (2010).
27. Lobo, T. J., Lansdorp, P. M. & Guryev, V. Local G-quadruplexes are not a major determinant of altered gene expression in BLM-deficient cells. Preprint at https://doi.org/10.1101/2023.09.08.556664 (2023).
28. Hickson, I. D. RecQ helicases: caretakers of the genome. *Nat Rev Cancer* **3**, 169–178 (2003).
29. Karow, J. K., Chakraverty, R. K. & Hickson, I. D. The Bloom's Syndrome Gene Product Is a 3′-5′ DNA Helicase. *J. Biol. Chem.* **272**, 30611–30614 (1997).
30. Popuri, V. *et al.* The Human RecQ helicases, BLM and RECQ1, display distinct DNA substrate specificities. *J. Biol. Chem.* **283**, 17766–17776 (2008).
31. Mohaghegh, P., Karow, J. K., Brosh, R. M., Bohr, V. A. & Hickson, I. D. The Bloom's and Werner's syndrome proteins are DNA structure-specific helicases. *Nucleic Acids Res.* **29**, 2843–2849 (2001).
32. van Brabant, A. J. *et al.* Binding and melting of D-loops by the Bloom syndrome helicase. *Biochemistry* **39**, 14617–14625 (2000).
33. Mendoza, O., Bourdoncle, A., Boulé, J.-B., Brosh, R. M. & Mergny, J.-L. G-quadruplexes and helicases. *Nucleic Acids Res.* **44**, 1989–2006 (2016).
34. Rhodes, D. & Lipps, H. J. G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res* **43**, 8627–8637 (2015).
35. Vindigni, A. & Hickson, I. D. RecQ helicases: multiple structures for multiple functions? *HFSP Journal* **3**, 12 (2009).
36. Kitano, K. Structural mechanisms of human RecQ helicases WRN and BLM. *Front. Genet.* **5**, (2014).
37. Shi, J. *et al.* A helical bundle in the N-terminal domain of the BLM helicase mediates dimer and potentially hexamer formation. *J. Biol. Chem.* **292**, 5909–5920 (2017).
38. Böhm, S. & Bernstein, K. A. The role of post-translational modifications in fine-tuning BLM helicase function during DNA repair. *DNA Repair* **22**, 123–132 (2014).
39. Bythell-Douglas, R. & Deans, A. J. A Structural Guide to the Bloom Syndrome Complex. *Structure* **29**, 99–113 (2021).

40. Gorbalenya, A. E. & Koonin, E. V. Helicases: amino acid sequence comparisons and structure-function relationships. *Current Opinion in Structural Biology* **3**, 419–429 (1993).
41. Killoran, M. P. & Keck, J. L. Sit down, relax and unwind: structural insights into RecQ helicase mechanisms. *Nucleic Acids Research* **34**, 4098–4105 (2006).
42. Bernstein, D. A. Domain mapping of Escherichia coli RecQ defines the roles of conserved N- and C-terminal regions in the RecQ family. *Nucleic Acids Research* **31**, 2778–2785 (2003).
43. Bahr, A., De Graeve, F., Kedinger, C. & Chatton, B. Point mutations causing Bloom's syndrome abolish ATPase and DNA helicase activities of the BLM protein. *Oncogene* **17**, 2565–2571 (1998).
44. German, J., Sanz, M. M., Ciocci, S., Ye, T. Z. & Ellis, N. A. Syndrome-causing mutations of the *BLM* gene in persons in the Bloom's Syndrome Registry. *Hum. Mutat.* **28**, 743–753 (2007).
45. Huber, M. D., Duquette, M. L., Shiels, J. C. & Maizels, N. A conserved G4 DNA binding domain in RecQ family helicases. *J. Mol. Biol.* **358**, 1071–1080 (2006).
46. Kitano, K., Kim, S.-Y. & Hakoshima, T. Structural Basis for DNA Strand Separation by the Unconventional Winged-Helix Domain of RecQ Helicase WRN. *Structure* **18**, 177–187 (2010).
47. Kim, Y. M. & Choi, B.-S. Structure and function of the regulatory HRDC domain from human Bloom syndrome protein. *Nucleic Acids Research* **38**, 7764–7777 (2010).
48. Kaneko, H. *et al.* Expression of the BLM gene in human haematopoietic cells. *Clin. Exp. Immunol.* **118**, 285–289 (1999).
49. Moens, P. B., Freire, R., Tarsounas, M., Spyropoulos, B. & Jackson, S. P. Expression and nuclear localization of BLM, a chromosome stability protein mutated in Bloom's syndrome, suggest a role in recombination during meiotic prophase. *J. Cell. Sci.* **113 ( Pt 4)**, 663–672 (2000).
50. Yankiwski, V., Marciniak, R. A., Guarente, L. & Neff, N. F. Nuclear structure in normal and Bloom syndrome cells. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5214–5219 (2000).
51. Hayakawa, S. *et al.* Characterization of the nuclear localization signal in the DNA helicase responsible for Bloom syndrome. *Int J Mol Med* (2000) doi:10.3892/ijmm.5.5.477.
52. Kaneko, H. *et al.* BLM (the Causative Gene of Bloom Syndrome) Protein Translocation into the Nucleus by a Nuclear Localization Signal. *Biochemical and Biophysical Research Communications* **240**, 348–353 (1997).
53. Davies, S. L., North, P. S. & Hickson, I. D. Role for BLM in replication-fork restart and suppression of origin firing after replicative stress. *Nat Struct Mol Biol* **14**, 677–679 (2007).
54. Fouché, N., Ozgür, S., Roy, D. & Griffith, J. D. Replication fork regression in repetitive DNAs. *Nucleic Acids Res.* **34**, 6044–6050 (2006).
55. Karow, J. K., Constantinou, A., Li, J.-L., West, S. C. & Hickson, I. D. The Bloom's syndrome gene product promotes branch migration of Holliday junctions. *Proceedings of the National Academy of Sciences* **97**, 6504–6508 (2000).
56. Machwe, A., Karale, R., Xu, X., Liu, Y. & Orren, D. K. The Werner and Bloom syndrome proteins help resolve replication blockage by converting (regressed) holliday junctions to functional replication forks. *Biochemistry* **50**, 6774–6788 (2011).

57. Syeda, A. H., Hawkins, M. & McGlynn, P. Recombination and Replication. *Cold Spring Harbor Perspectives in Biology* **6**, a016550–a016550 (2014).
58. Wilson, D. M. & Thompson, L. H. Molecular mechanisms of sister-chromatid exchange. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **616**, 11–23 (2007).
59. Wu, L. Role of the BLM helicase in replication fork management. *DNA Repair* **6**, 936–944 (2007).
60. Cesare, A. J. & Reddel, R. R. Alternative lengthening of telomeres: models, mechanisms and implications. *Nat Rev Genet* **11**, 319–330 (2010).
61. Du, X. *et al.* Telomere shortening exposes functions for the mouse Werner and Bloom syndrome genes. *Mol. Cell. Biol.* **24**, 8437–8446 (2004).
62. Lillard-Wetherell, K. *et al.* Association and regulation of the BLM helicase by the telomere proteins TRF1 and TRF2. *Hum. Mol. Genet.* **13**, 1919–1932 (2004).
63. Mendez-Bermudez, A. *et al.* The roles of WRN and BLM RecQ helicases in the Alternative Lengthening of Telomeres. *Nucleic Acids Res.* **40**, 10809–10820 (2012).
64. Opresko, P. L. *et al.* POT1 stimulates RecQ helicases WRN and BLM to unwind telomeric DNA substrates. *J. Biol. Chem.* **280**, 32069–32080 (2005).
65. Pan, X. *et al.* FANCM, BRCA1, and BLM cooperatively resolve the replication stress at the ALT telomeres. *Proc Natl Acad Sci USA* **114**, E5940–E5949 (2017).
66. Schawalder, J., Paric, E. & Neff, N. F. Telomere and ribosomal DNA repeats are chromosomal targets of the bloom syndrome DNA helicase. *BMC Cell Biol.* **4**, 15 (2003).
67. Bischof, O. *et al.* Regulation and localization of the Bloom syndrome protein in response to DNA damage. *J. Cell Biol.* **153**, 367–380 (2001).
68. Croteau, D. L., Popuri, V., Opresko, P. L. & Bohr, V. A. Human RecQ Helicases in DNA Repair, Recombination, and Replication. *Annu. Rev. Biochem.* **83**, 519–552 (2014).
69. Carney, J. P. *et al.* The hMre11/hRad50 protein complex and Nijmegen breakage syndrome: linkage of double-strand break repair to the cellular DNA damage response. *Cell* **93**, 477–486 (1998).
70. Nimonkar, A. V. *et al.* 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 (2011).
71. Trujillo, K. M., Yuan, S. S., Lee, E. Y. & Sung, P. Nuclease activities in a complex of human recombination and DNA repair factors Rad50, Mre11, and p95. *J. Biol. Chem.* **273**, 21447–21450 (1998).
72. Wright, W. D., Shah, S. S. & Heyer, W.-D. Homologous recombination and the repair of DNA double-strand breaks. *J. Biol. Chem.* **293**, 10524–10535 (2018).
73. Manthei, K. A. & Keck, J. L. The BLM dissolvasome in DNA replication and repair. *Cell. Mol. Life Sci.* **70**, 4067–4084 (2013).
74. Krejci, L., Altmannova, V., Spirek, M. & Zhao, X. Homologous recombination and its regulation. 24.
75. Kowalczykowski, S. C. An Overview of the Molecular Mechanisms of Recombinational DNA Repair. 38.
76. Roth, N. *et al.* The requirement for recombination factors differs considerably between different pathways of homologous double-strand break repair in somatic plant cells. *The Plant Journal* **72**, 781–790 (2012).

77. Shah Punatar, R., Martin, M. J., Wyatt, H. D. M., Chan, Y. W. & West, S. C. Resolution of single and double Holliday junction recombination intermediates by GEN1. *Proc Natl Acad Sci USA* **114**, 443–450 (2017).
78. Wyatt, H. D. M. & West, S. C. Holliday Junction Resolvases. *Cold Spring Harbor Perspectives in Biology* **6**, a023192–a023192 (2014).
79. Carroll, D. Genetic Recombination. in *Encyclopedia of Genetics* 841–845 (Elsevier, 2001). doi:10.1006/rwgn.2001.0543.
80. Rédei, G. P. *Encyclopedia of Genetics, Genomics, Proteomics, and Informatics*. (Springer, Berlin, 2008).
81. Bugreev, D. V., Yu, X., Egelman, E. H. & Mazin, A. V. Novel pro- and anti-recombination activities of the Bloom's syndrome helicase. *Genes Dev.* **21**, 3085–3094 (2007).
82. Baudat, F. Meiotic recombination in mammals: localization and regulation. *g e n e ti c s* 13 (2013).
83. Lisby, M. & Rothstein, R. Cell Biology of Mitotic Recombination. *Cold Spring Harb Perspect Biol* **7**, a016535 (2015).
84. Chaganti, R. S. K., Schonberg, S. & German, J. A Manyfold Increase in Sister Chromatid Exchanges in Bloom's Syndrome Lymphocytes. *Proc. Natl. Acad. Sci. U.S.A.* **71**, 4508–4512 (1974).
85. Holloway, J. K., Morelli, M. A., Borst, P. L. & Cohen, P. E. Mammalian BLM helicase is critical for integrating multiple pathways of meiotic recombination. *Journal of Cell Biology* **188**, 779–789 (2010).
86. Annus, T. *et al.* Bloom syndrome helicase contributes to germ line development and longevity in zebrafish. *Cell Death Dis* **13**, 363 (2022).
87. Hatkevich, T. *et al.* Bloom Syndrome Helicase Promotes Meiotic Crossover Patterning and Homolog Disjunction. *Current Biology* **27**, 96–102 (2017).
88. Lazzarano, S. *et al.* Genetic mapping of species differences via in vitro crosses in mouse embryonic stem cells. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3680–3685 (2018).
89. Song, J. H. T. *et al.* Genetic studies of human–chimpanzee divergence using stem cell fusions. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2117557118 (2021).
90. German, J. Bloom's syndrome. I. Genetical and clinical observations in the first twenty-seven patients. *Am. J. Hum. Genet.* **21**, 196–227 (1969).
91. Li, L., Eng, C., Desnick, R. J., German, J. & Ellis, N. A. Carrier Frequency of the Bloom SyndromeblmAshMutation in the Ashkenazi Jewish Population. *Molecular Genetics and Metabolism* **64**, 286–290 (1998).
92. Neff, N. F. *et al.* The DNA Helicase Activity of BLM Is Necessary for the Correction of the Genomic Instability of Bloom Syndrome Cells. *MBoC* **10**, 665–676 (1999).
93. Rong, S. B., Väliaho, J. & Vihinen, M. Structural basis of Bloom syndrome (BS) causing mutations in the BLM helicase domain. *Mol. Med.* **6**, 155–164 (2000).
94. Neff, N. F. *et al.* The DNA Helicase Activity of BLM Is Necessary for the Correction of the Genomic Instability of Bloom Syndrome Cells. *MBoC* **10**, 665–676 (1999).
95. Gruber, S. B. BLM Heterozygosity and the Risk of Colorectal Cancer. *Science* **297**, 2013–2013 (2002).
96. Alzahrani, F. A. *et al.* Investigating the pathogenic SNPs in BLM helicase and their biological consequences by computational approach. *Sci Rep* **10**, 12377 (2020).
97. De Renty, C. & Ellis, N. A. Bloom's syndrome: Why not premature aging? *Ageing Research Reviews* **33**, 36–51 (2017).

98. German, J. Bloom's syndrome. *Dermatol Clin* **13**, 7–18 (1995).
99. Cunniff, C., Bassetti, J. A. & Ellis, N. A. Bloom's Syndrome: Clinical Spectrum, Molecular Pathogenesis, and Cancer Predisposition. *Mol Syndromol* **8**, 4–23 (2017).
100. German, J. Bloom's syndrome. XX. The first 100 cancers. *Cancer Genetics and Cytogenetics* **93**, 100–106 (1997).
101. Arora, H. *et al.* Bloom syndrome. *Int J Dermatol* **53**, 798–802 (2014).
102. Gretzula, J. C., Hevia, O. & Weber, P. J. Bloom's syndrome. *Journal of the American Academy of Dermatology* **17**, 479–488 (1987).
103. Kauli, R., Prager-Lewin, R., Kaufman, H. & Laron, Z. GONADAL FUNCTION IN BLOOM'S SYNDROME. *Clinical Endocrinology* **6**, 285–289 (1977).
104. Shabtai, F. & Halbrecht, I. Bloom's syndrome, missing Y, hypogonadism and cancer. *Clinical Genetics* **18**, 93–95 (1980).
105. Amor-Gueret, M. Bloom Syndrome. in *Encyclopedia of Cancer* (ed. Schwab, M.) 438–440 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011). doi:10.1007/978-3-642-16483-5_671.
106. German, J., Crippa, L. P. & Bloom, D. Bloom's syndrome. III. Analysis of the chromosome aberration characteristic of this disorder. *Chromosoma* **48**, 361–366 (1974).
107. Gaymes, T. J. *et al.* Increased error-prone non homologous DNA end-joining--a proposed mechanism of chromosomal instability in Bloom's syndrome. *Oncogene* **21**, 2525–2533 (2002).
108. Kinzler, K. W. & Vogelstein, B. Lessons from Hereditary Colorectal Cancer. *Cell* **87**, 159–170 (1996).
109. Knudson, A. G. Antioncogenes and human cancer. *PNAS* **90**, 10914–10921 (1993).
110. van Wietmarschen, N. *et al.* BLM helicase suppresses recombination at G-quadruplex motifs in transcribed genes. *Nat Commun* **9**, 271 (2018).
111. Rosin, M. P. & German, J. Evidence for chromosome instability in vivo in bloom syndrome: Increased numbers of micronuclei in exfoliated cells. *Hum Genet* **71**, 187–191 (1985).
112. Frorath, B., Schmidt-Preuss, U., Z llner, M. & R diger, H. W. Heterozygous carriers for Bloom syndrome exhibit a spontaneously increased micronucleus formation in cultured fibroblasts. *Hum Genet* **67**, 52–55 (1984).
113. Krupina, K., Goginashvili, A. & Cleveland, D. W. Causes and consequences of micronuclei. *Current Opinion in Cell Biology* **70**, 91–99 (2021).
114. Luzhna, L., Kathiria, P. & Kovalchuk, O. Micronuclei in genotoxicity assessment: from genetics to epigenetics and beyond. *Front. Genet.* **4**, (2013).
115. Henderson, E., Hardin, C. C., Walk, S. K., Tinoco, I. & Blackburn, E. H. Telomeric DNA oligonucleotides form novel intramolecular structures containing guanine·guanine base pairs. *Cell* **51**, 899–908 (1987).
116. Dell'Oca, M. C. *et al.* Spotlight on G-Quadruplexes: From Structure and Modulation to Physiological and Pathological Roles. *IJMS* **25**, 3162 (2024).
117. Hänsel-Hertsch, R., Di Antonio, M. & Balasubramanian, S. DNA G-quadruplexes in the human genome: detection, functions and therapeutic potential. *Nat Rev Mol Cell Biol* **18**, 279–284 (2017).

118. Javadekar, S. M., Nilavar, N. M., Paranjape, A., Das, K. & Raghavan, S. C. Characterization of G-quadruplex antibody reveals differential specificity for G4 DNA forms. *DNA Res* **27**, dsaa024 (2020).
119. Esnault, C. *et al.* G4access identifies G-quadruplexes and their associations with open chromatin and imprinting control regions. *Nat Genet* **55**, 1359–1369 (2023).
120. Mendoza, O., Bourdoncle, A., Boulé, J.-B., Brosh, R. M. & Mergny, J.-L. G-quadruplexes and helicases. *Nucleic Acids Res* **44**, 1989–2006 (2016).
121. Spiegel, J. *et al.* G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biol* **22**, 117 (2021).
122. Zhang, X., Spiegel, J., Martínez Cuesta, S., Adhikari, S. & Balasubramanian, S. Chemical profiling of DNA G-quadruplex-interacting proteins in live cells. *Nat. Chem.* **13**, 626–633 (2021).
123. Biffi, G., Tannahill, D., McCafferty, J. & Balasubramanian, S. Quantitative visualization of DNA G-quadruplex structures in human cells. *Nature Chem* **5**, 182–186 (2013).
124. Henderson, A. *et al.* Detection of G-quadruplex DNA in mammalian cells. *Nucleic Acids Res* **42**, 860–869 (2014).
125. Wang, Y.-H. *et al.* G4LDB 2.2: a database for discovering and studying G-quadruplex and i-Motif ligands. *Nucleic Acids Research* **50**, D150–D160 (2022).
126. Balasubramanian, S., Hurley, L. H. & Neidle, S. Targeting G-quadruplexes in gene promoters: a novel anticancer strategy? *Nat Rev Drug Discov* **10**, 261–275 (2011).
127. Varshney, D., Spiegel, J., Zyner, K., Tannahill, D. & Balasubramanian, S. The regulation and functions of DNA and RNA G-quadruplexes. *Nat Rev Mol Cell Biol* **21**, 459–474 (2020).
128. Kaya-Okur, H. S. *et al.* CUT&Tag for efficient epigenomic profiling of small samples and single cells. *Nat Commun* **10**, 1930 (2019).
129. Lee, W. T. C. *et al.* Single-molecule imaging reveals replication fork coupled formation of G-quadruplex structures hinders local replication stress signaling. *Nat Commun* **12**, 2525 (2021).
130. Cayrou, C. *et al.* Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res* **21**, 1438–1449 (2011).
131. Cayrou, C. *et al.* The chromatin environment shapes DNA replication origin organization and defines origin classes. *Genome Res* **25**, 1873–1885 (2015).
132. Comoglio, F. *et al.* High-resolution profiling of Drosophila replication start sites reveals a DNA shape and chromatin signature of metazoan origins. *Cell Rep* **11**, 821–834 (2015).
133. Langley, A. R., Gräf, S., Smith, J. C. & Krude, T. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). *Nucleic Acids Res* **44**, 10230–10247 (2016).
134. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat Commun* **10**, 3274 (2019).
135. Meyne, J., Ratliff, R. L. & Moyzis, R. K. Conservation of the human telomere sequence (TTAGGG)n among vertebrates. *Proc Natl Acad Sci U S A* **86**, 7049–7053 (1989).
136. Dumetz, F. & Merrick, C. Parasitic Protozoa: Unusual Roles for G-Quadruplexes in Early-Diverging Eukaryotes. *Molecules* **24**, 1339 (2019).

137. Oganesian, L., Moon, I. K., Bryan, T. M. & Jarstfer, M. B. Extension of G-quadruplex DNA by ciliate telomerase. *EMBO J* **25**, 1148–1159 (2006).
138. Crabbe, L., Verdun, R. E., Haggblom, C. I. & Karlseder, J. Defective Telomere Lagging Strand Synthesis in Cells Lacking WRN Helicase Activity. *Science* **306**, 1951–1953 (2004).
139. Drosopoulos, W. C., Kosiyatrakul, S. T. & Schildkraut, C. L. BLM helicase facilitates telomere replication during leading strand synthesis of telomeres. *Journal of Cell Biology* **210**, 191–208 (2015).
140. Vannier, J.-B., Pavicic-Kaltenbrunner, V., Petalcorin, M. I. R., Ding, H. & Boulton, S. J. RTEL1 Dismantles T Loops and Counteracts Telomeric G4-DNA to Maintain Telomere Integrity. *Cell* **149**, 795–806 (2012).
141. Kanoh, Y. *et al.* Rif1 binds to G quadruplexes and suppresses replication over long distances. *Nat Struct Mol Biol* **22**, 889–897 (2015).
142. Biffi, G., Tannahill, D. & Balasubramanian, S. An Intramolecular G-Quadruplex Structure Is Required for Binding of Telomeric Repeat-Containing RNA to the Telomeric Protein TRF2. *J. Am. Chem. Soc.* **134**, 11974–11976 (2012).
143. Siddiqui-Jain, A., Grand, C. L., Bearss, D. J. & Hurley, L. H. Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-*MYC* transcription. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11593–11598 (2002).
144. Esain-Garcia, I. *et al.* G-quadruplex DNA structure is a positive regulator of *MYC* transcription. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2320240121 (2024).
145. Li, C. *et al.* Ligand-induced native G-quadruplex stabilization impairs transcription initiation. *Genome Res.* **31**, 1546–1560 (2021).
146. Wulfridge, P. *et al.* G-quadruplexes associated with R-loops promote CTCF binding. *Molecular Cell* **83**, 3064-3079.e5 (2023).
147. Lee, C.-Y. *et al.* R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. *Nat Commun* **11**, 3392 (2020).
148. Raiber, E.-A., Kranaster, R., Lam, E., Nikan, M. & Balasubramanian, S. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Res* **40**, 1499–1508 (2012).
149. Mao, S.-Q. *et al.* DNA G-quadruplex structures mold the DNA methylome. *Nat Struct Mol Biol* **25**, 951–957 (2018).
150. Falconer, E. *et al.* DNA template strand sequencing of single-cells maps genomic rearrangements at high resolution. *Nat Methods* **9**, 1107–1112 (2012).
151. Castillo Bosch, P. *et al.* FANCJ promotes DNA synthesis through G-quadruplex structures. *EMBO J* **33**, 2521–2533 (2014).
152. Schult, P., Simon, P. & Paeschke, K. G-Quadruplex Resolving by Specific Helicases. in *Handbook of Chemical Biology of Nucleic Acids* (ed. Sugimoto, N.) 1–18 (Springer Nature Singapore, Singapore, 2023). doi:10.1007/978-981-16-1313-5_101-1.
153. Nguyen, G. H. *et al.* A Small Molecule Inhibitor of the BLM Helicase Modulates Chromosome Stability in Human Cells. *Chemistry & Biology* **20**, 55–62 (2013).
154. Liu, S. *et al.* Aneuploidy effects on human gene expression across three cell types. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218478120 (2023).
155. Atri, A. The Alzheimer's Disease Clinical Spectrum: Diagnosis and Management. *Med Clin North Am* **103**, 263–293 (2019).

156. Ge, Y., Paisie, T. K., Chen, S. & Concannon, P. UBASH3A Regulates the Synthesis and Dynamics of TCR–CD3 Complexes. *The Journal of Immunology* **203**, 2827–2836 (2019).
157. Mitteaux, J. *et al.* Identifying G-Quadruplex-DNA-Disrupting Small Molecules. *J. Am. Chem. Soc.* **143**, 12567–12577 (2021).
158. Neupane, A., Chariker, J. H. & Rouchka, E. C. Analysis of nucleotide variations in human g-quadruplex forming regions associated with disease states. Preprint at https://doi.org/10.1101/2023.01.30.526341 (2023).

# Appendix

# Appendix 1

## Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Volker Soltys[1]*, Moritz Peters[1], Dingwen Su[1], Marek Kučka[1,2], Yingguang Frank Chan[1,3]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany 2 Department of Translational Genomics, University of Cologne, 50931 Cologne, Germany 3 University of Groningen, Groningen Institute of Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

* Corresponding authors volker.soltys@tue.mpg.de; frank.chan@rug.nl

# Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells

Volker Soltys[1*], Moritz Peters[1], Dingwen Su[1], Marek Kučka[1,2,], Yingguang Frank Chan[1,3]

1 Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany
2 Department of Translational Genomics, University of Cologne, 50931 Cologne, Germany
3 University of Groningen, Groningen Institute of Evolutionary Life Sciences, 9747 AG Groningen, Netherlands

* Corresponding authors
volker.soltys@tue.mpg.de; frank.chan@rug.nl

## Abstract

Gene expression and chromatin accessibility are highly interconnected processes. Disentangling one without the other provides an incomplete picture of gene regulation. However, simultaneous measurements of RNA and accessible chromatin are technically challenging, especially when studying complex organs with rare cell-types. Here, we present easySHARE-seq, an elaboration of SHARE-seq, providing simultaneous measurements of ATAC- and RNA-seq within single cells, enabling identification of cell-type specific *cis*-regulatory elements (CREs). easySHARE-seq retains high scalability, improves RNA-seq data quality while also allowing for flexible study design. Using 19,664 joint profiles from murine liver nuclei, we linked CREs to their target genes and uncovered complex regulation of key genes such as *Gata4*. We further identify *de-novo* genes and *cis*-regulatory elements displaying zonation in Liver sinusoidal epithelial cells (LSECs), a challenging cell type with low mRNA levels, demonstrating the power of multimodal measurements. EasySHARE-seq therefore provides a flexible platform for investigating gene regulation across cell types and scale.

# *Appendix 1*

*Flexible and high-throughput simultaneous profiling of gene expression and chromatin accessibility in single cells*



## Introduction

Gene expression and chromatin state combined influence fundamental processes such as gene regulation or cell fate decisions [1–3]. A better understanding of these mechanisms and their interactions will be a major step toward decoding developmental trajectories or reconstructing cellular taxonomies in both health and disease. However, to fully capture these complex relationships, multiple information layers need to be measured simultaneously. For example, prior studies have argued that chromatin state is often predictive of gene expression and can also prime cells toward certain lineage decisions or even induce tissue regeneration[4–6]. However, these studies depend on the computational integration of separately measured modalities. By assuming a shared biological state, this restricts the discovery of novel and potentially fine-scale differences and renders it challenging to identify the root cause of erroneous cell states[7].

The last decade has seen an explosive growth in single-cell methodologies, with new assays, increasing throughput and a suite of computational tools[8]. Most non-commercial high-throughput methodologies rely on combinatorial indexing for single-cell barcoding, where sequential rounds of barcodes combine to create unique cellular barcode combinations [9,10]. Compared to single-modality assays, multi-omic technologies, which capture two or more information layers, are relatively new. Therefore, they are still limited in sensitivity and throughput and commercial kits can be expensive such that multi-omic studies tend to have limited sample sizes [11,12].

To address these problems, we built upon a previously developed protocol called SHARE-seq[13] and developed easySHARE-seq, a protocol for simultaneously measuring gene expression and chromatin accessibility within single cells using combinatorial indexing. Major improvements include easySHARE-seq's barcoding framework, which allows for expanded and flexible study design, all while being compatible with standard Illumina sequencing, thereby removing economic hurdles. Importantly, easySHARE-seq retains the scalability and improves upon RNA-seq sensitivity of the original SHARE-seq protocol. Here, we used easySHARE-seq to profile 19,664 murine liver nuclei and show that we can recover high quality data in both RNA-seq and ATAC-seq channels, which are highly congruent and share equal power in classifying cell types. We then surveyed the *cis*-regulatory landscape of Liver Sinusoidal Endothelial Cells (LSECs), leveraging the simultaneous measurements of gene expression and chromatin accessibility and identified 40,957 links between expressed genes and nearby ATAC-seq peaks. Notably, genes with the highest number of links were enriched for transcription factors and regulators known to control important functions within LSECs. Lastly, we show that easySHARE-seq can be used to investigate micro-scale changes in accessibility and gene expression by identifying novel markers and open chromatin regions displaying zonation in LSECs. This technology improves our toolkit of multi-omic protocols needed for advancing our knowledge about gene regulation and cell fate decisions.

## Results

### easySHARE-seq reliably labels both transcriptome and accessible chromatin in individual cells

To develop a multi-omic single-cell (sc) RNA and scATAC-seq protocol that allows for flexible study design while being highly scalable, we built upon SHARE-seq[13] to create easySHARE-seq, which uses two rounds of ligation to simultaneously label cDNA and DNA fragments in the same cell (**Fig. 1A**). Due to a much more streamlined barcoding structure, easySHARE-seq allows 300bp sequencing of the insert. This longer read-length leads to a higher recovery of DNA variants, thus increasing the power to detect allele-specific signals or cell-specific variation, e.g., in hybrids or cancer cells[14].

To generate libraries, fixed and permeabilized cells or nuclei (we will use "cells" afterwards to refer to both) are transposed by Tn5 transposase carrying a custom adapter with a single-stranded overhang (**Fig. 1B**). Next, mRNA is reverse transcribed (RT) using a biotinylated poly(T) primer with an identical overhang. Subsequently, the cells are individually barcoded in two rounds of combinatorial indexing with 192 barcodes in each round, creating a total of 36,864 possible barcode combinations.The first barcode is ligated onto the already present overhang and itself contains a second single-stranded overhang, onto which the second barcode is ligated to. Importantly, in the easySHARE-seq design, we have kept the total length of the barcode within 17nt ("Index 1" read; **Fig. 1B, Suppl. Fig. 1A**), allowing for multiplexing of easySHARE-seq libraries with standard Illumina libraries. In contrast, in the original publication, SHARE-seq libraries required Index 1 lengths of 99nt, a highly custom configuration which would require a costly private sequencing run most of the time.

After barcoding, the cells are aliquoted into sub-libraries of approximately 3,500 cells each and reverse crosslinked. A streptavidin pull-down of the biotinylated RT-primer is performed to separate the cDNA molecules from the chromatin ("fragments"). Each sub-library is then prepared for sequencing and amplified using matched indexing primers to allow identification of paired cellular scRNA- and scATAC-seq profiles. By scaling up the numbers of sub-libraries, this barcoding strategy therefore allows for high-throughput experiments of hundreds of thousands of cells, only limited by the availability of indexing primers. For a detailed description of the flexibility of easySHARE-seq, instructions on how to modify and incorporate the framework into new designs as well as critical steps to assess when planning to use easySHARE-seq see Supplementary Notes.

To evaluate the accuracy and cell-specificity of the barcoding, we first performed easySHARE-seq on a mixed pool between human and murine cell lines (HEK and OP-9). This design allows us to identify two or more cells sharing the same barcode ('doublets'; **Fig. 1C**, left). After sequencing, we recovered a total of 3,808 cells. Both chromatin and transcriptome profiles separated well within each cell (**Fig. 1C**, middle), with cDNA showing a lower accuracy with increasing transcript counts, likely due to less precise read mapping. We identified a total of 124 doublets (**Fig, 1C**, right), which gives a final doublet rate of 6.34% factoring in the undetectable intra-species doublets. For comparison, a 10X Chromium Next GEM experiment with 10,000 cells has a doublet rate of ~7.9% (www.10xgenomics.com). Importantly, easySHARE-seq doublet rates can be lowered further by aliquoting fewer cells within each sub-library. To summarise, easySHARE-seq provides a high-throughput and flexibility framework for accurately measuring chromatin accessibility and gene expression in single cells.

**Simultaneous scATAC-seq and scRNA-seq profiling in murine primary liver cells**

To assess data quality and investigate the relationship between gene expression and chromatin accessibility, we focused on murine liver. The liver consists of a diverse set of defined primary cell types, ranging from large and potentially multinucleated hepatocytes to small non-parenchymal cell types such as Liver Sinusoidal Endothelial Cells[15] (LSECs).

We generated matched high-quality chromatin and gene expression profiles for 19,664 adult liver cells across four age-matched mice (2 male, 2 female), amounting to a recovery rate of 70.2% (28,000 input cells). Each nuclei had on average 3,629 UMIs and 2,213 fragments (74% of all RNA-seq reads were cDNA, 55.9% mean ATAC-seq fragments in peaks; **Suppl. Fig. 1B & D**). In terms of UMIs per cell, easySHARE-seq therefore out-performed other previously published multi-omic and representative single channel assays (**Fig. 2B**; see figure legend for tissue type and study). Consistent with nuclei as input material, the majority of cDNA molecules were intronic (69.6%, **Suppl. Fig. 1C & H**). Regarding DNA fragments per cell, easySHARE-seq performed similarly to other published multi-omic assays (**Fig. 2C**) and scATAC-seq libraries displayed the characteristic banding pattern with reads being highly enriched at transcription start sites (TSS; **Suppl. Fig. 1 E, F, H**).

To visualise and identify cell types, we first projected the ATAC- and RNA-seq modalities separately into 2D Space and then used Weighted Nearest Neighbor[16] (WNN) integration to combine both modalities into a single UMAP visualisation (**Fig. 2A)**. Importantly, the same cells independently clustered together in the scRNA- and scATAC-seq UMAPs, showcasing high congruence between the two modalities (**Suppl. Fig. 2A&B**). We then annotated previously published cell types based on gene expression of previously established marker genes [17,18]. Marker gene expression was highly specific to the clusters (**Fig. 2D, Suppl. Fig. 2F**) and we recovered all expected cell types (**Suppl. Fig.2C**). Importantly, the same cell types were identified using each modality independently, showcasing high congruence between the scATAC- and scRNA-seq modalities (**Fig. 2E**). Altogether, our results show that easySHARE-seq generates high quality joint cellular profiles of chromatin accessibility and gene expression within primary tissue, expanding our toolkit of multi-omic protocols.

**Uncovering the cis-regulatory landscape of key regulators through peak-gene associations**

As easySHARE-seq simultaneously measures chromatin accessibility and gene expression, it allows to directly investigate the relationship between them to potentially connect *cis*-regulatory elements (CREs) to their target genes. To do so, we adopted the analytical framework from Ma et al.[13], which queries if an increased expression within a cell is significantly correlated with chromatin accessibility at a peak while controlling for GC content and accessibility strength. Focusing on LSECs (1,501 cells), we calculated associations between putative CREs (pCREs, defined as peaks with a significant peak–gene association) and each expressed gene, considering all peaks within ± 500kb of the TSS. We identified 40,957 significant peak–gene associations (45% of total peaks, $P < 0.05$, FDR = 0.1) with 15,061 genes having at least one association (76.8% of all expressed genes, **Suppl. Fig. 3A,C**). In rare cases (2.9%), these pCREs were associated with five or more genes, which drop to 0.03% when considering only pCREs within ± 50 kb of a TSS (**Suppl. Fig. 3B,D**). These pCREs tended to cluster to regions of higher expressed gene density (2.15 mean expressed genes within 50kbp vs 0.93 for all global peaks) and their associated genes were enriched for biological processes such as mRNA processing, histone modifications and splicing (**Suppl. Fig. 3H**), possibly reflecting loci with increased regulatory activity.

We then ranked the genes based on their number of associated pCREs (**Fig. 2F**). Within the top 1% genes with the most pCRE associations were many key regulators and transcription factors. Examples include *Taf5,* which directly binds the TATA-box[19] and is required for initiation of transcription, or *Gata4,* which has been identified as the master regulator for LSEC specification during development as well as controlling regeneration and metabolic maturation of liver tissue in adult mice [20,21]. As such, it incorporates a variety of signals and its expression needs to be strictly regulated, which is reflected in its many pCREs associations (**Fig. 2H**). Similarly, *Igf1* also integrates signals from many different pCREs[22] (**Suppl. Fig. 3G**). Notably, pCRES are significantly enriched at transcription start sites (TSS), even relative to background enrichment (**Fig. 2G**).

To summarise, easySHARE-seq allows to directly investigate the relationship between chromatin accessibility and gene expression and identify putative *cis*-regulatory elements at genomic scale, even in small cell types with relatively low mRNA contents (**Suppl. Fig. 2D)**.

***De Novo* identification of open chromatin regions and genes displaying zonation in LSECs**

We next investigated the process of zonation in LSECs. The liver consists of hexagonal units called lobules where blood flows from the portal vein and arteries toward a central vein [23,24] (**Fig. 3A)**. The central–portal (CP) axis is characterised by a morphogen gradient, e.g. from *Wnt2*, secreted by central vein LSECs, with the resulting micro-environment giving rise to spatial division of labour among hepatocytes [25–27]. Studying zonation in non-parenchymal cells such as LSECs is challenging as these are small cells with low mRNA content (**Suppl. Fig. 2D,E**), lying below the detection limit of current spatial transcriptomic techniques. As a result, only very few studies assess zonation in LSECs on a genomic level [28]. However, LSECs are critical to liver function as they line the artery walls, clear and process endotoxins, play a critical role in liver regeneration and secrete morphogens themselves to regulate hepatocyte gene expression [29–31], rendering their understanding a prerequisite for tackling many diseases.

We therefore asked if we can recover known zonation gradients and potentially identify novel marker genes and open chromatin regions displaying zonation. We noticed that LSECs clustered in a distinct linear pattern in our UMAP projection and therefore divided them into equal bins along UMAP2 coordinates (**Suppl. Fig. 4A**, number of cells per bin 80-260, median: 128). We then calculated mean normalised expression and mean normalised accessibility within each bin. This recovered gene expression and chromatin accessibility gradients for major known zonation marker genes[28] (**Fig. 3B,C**). For example, *Wnt2* expression decreased strongly along the CP axis as did chromatin accessibility of all three peaks at the *Wnt2* locus (**Fig. 3B**). We also recovered the zonation profiles for the majority of known pericentral (increasing along the CP-axis), periportal (decrease along the CP-axis) and non-monotonic markers (decrease toward both ends) as well as their associated chromatin regions (**Fig. 3C**). Gene expression zonation profiles can also be recovered by ordering LSECs along pseudotime (**Suppl. Fig. 4C,D**). In contrast, simply subclustering LSECs and comparing expression between these clusters was too broad for the assessment of zonation (**Suppl. Fig. 4A,B**).

Next, we sought to identify novel marker genes and open chromatin regions displaying zonation in LSECs based on the decrease or increase of mean expression or accessibility along the previously established bins. In total, we classified 153 genes and 381 open chromatin regions as pericentral and 209 genes and 465 open chromatin regions showed periportal zonation profiles (**Fig. 3D**). The list of markers contained many genes regulating epithelial growth and angiogenesis (e.g. *Efna1, Nrg2, Zfpm1, Zfpm2, Bmpr2*)[32–34], related to

regulating hepatocyte functions and communication (e.g. *Dll4, Foxo1, Sp1, Snx3*)[35–37] as well as immunological functions (e.g. *Sirt2, Cd59a)*[38,39]*,* suggesting that these processes show variation along the PC axis. As dysregulation of LSEC zonation is implicated in multiple illnesses such as liver cirrhosis or non-alcoholic fatty liver disease [40,41], these genes are potential new biomarkers for their identification and the open chromatin regions starting points for investigating the role of gene regulation in their emergence.

## Discussion

Understanding complex processes such as gene regulation or disease states requires the integration of multiple layers of information. Here, we show that easySHARE-seq provides a high-quality, high-throughput and flexible platform for joint profiling of chromatin accessibility and gene expression within single cells. We show that both modalities are highly congruent with one another and we leverage their simultaneous measurements to identify peak–gene interactions and survey the *cis*-regulatory landscape of LSECs. We also show that easySHARE-seq can be used to assess micro-scale changes such as zonation in LSECs across both gene expression and chromatin accessibility. These cells have low mRNA content and we recovered zonation profiles of many transcription factors, which are often lowly expressed, further demonstrating the power of easySHARE-seq.

Besides improving upon RNA-seq data quality, we argue that easySHARE-seq has many advantages, especially in terms of the sequencing flexibility due to the barcode design, which can help remove hurdles for incorporating multi-omic single-cell assays into study designs. Combined with shorter experimental times (~12h total), easySHARE-seq might be particularly suited for studies where higher sample sizes are required or ones that rely on identification of genomic variants, e.g., in diverse, non-inbred individuals or in cancer. In terms of costs per cell, easySHARE-seq performs similarly to standard SHARE-seq with ~0.056 cents/cell, a fraction of the costs (<25%) of commercially available platforms, even before factoring in the specialized instrument costs. A comparison between technologies can be found in **Table 1**.

We envision easySHARE-seq as another technological step toward ultimately understanding gene regulation in health and disease, surveying *cis*-regulatory landscapes during differentiation and lineage commitment and determining genetic variants affecting those processes.

**Figure 1**

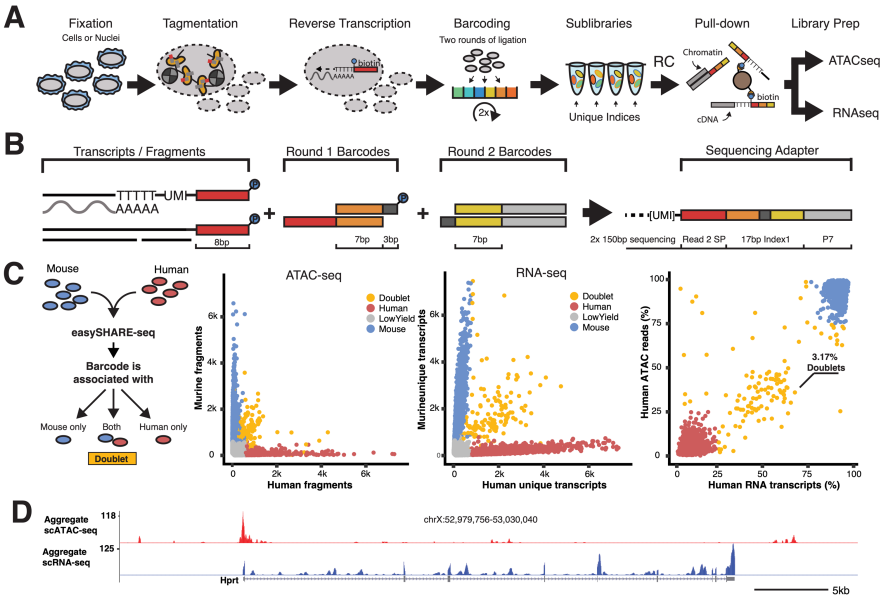



**Figure 1: easySHARE-seq enables highly-accurate simultaneous scATAC-seq and scRNA-seq profiling**

(A) Schematic workflow of easySHARE-seq.
(B) Generation and structure of the single-cell barcoding within Index 1.
(C) Principle of a species-mixing experiment. Cells are mixed prior to easySHARE-seq and sequences associated with each cell barcode are assessed for genome of origin (left panel). Unique ATAC fragments per cell aligning to the mouse or human genome (middle left). Cells are coloured according to their assigned origin (red: human; blue: mouse; orange: doublet). Middle right: Same plot but with RNA UMIs. Right: Percentage of ATAC fragments or RNA UMIs per cell relative to total sequencing reads mapping uniquely to the human genome. 3.17% of all observed cells classified as doublets. Accounting for same-species doublets, this results in a doublet rate of 6.34%.
(D) Aggregate chromatin accessibility (red) and expression-seq (blue) profile of OP-9 cells at the *Hprt* locus.

## Figure 2

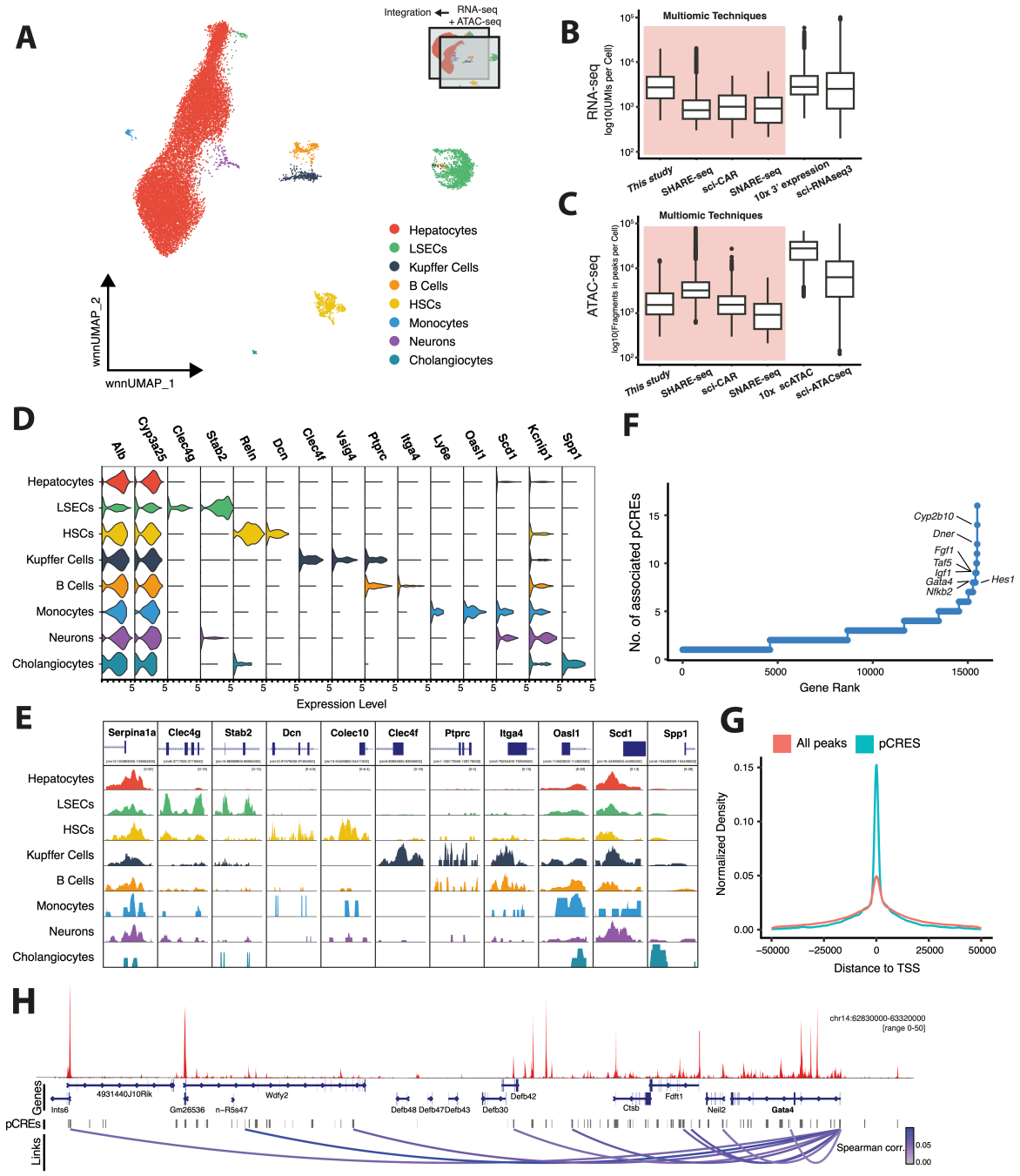

A
Integration
RNA-seq
+ ATAC-seq
Hepatocytes
LSECs
Kupffer Cells
B Cells
HSCs
Monocytes
Neurons
Cholangiocytes
wnnUMAP_2
wnnUMAP_1
B
Multiomic Techniques
RNA-seq
log10(UMIs per Cell)
This study
SHARE-seq
sci-CAR
SNARE-seq
10x 3' expression
sci-RNAseq3
C
Multiomic Techniques
ATAC-seq
log10(Fragments in peaks per Cell)
This study
SHARE-seq
sci-CAR
SNARE-seq
10x scATAC
sci-ATACseq
D
Alb
Cyp3a25
Clec4g
Stab2
Reln
Dcn
Clec4f
Vsig4
Ptprc
Itga4
Ly6e
Oasl1
Scd1
Kcnip1
Spp1
Hepatocytes
LSECs
HSCs
Kupffer Cells
B Cells
Monocytes
Neurons
Cholangiocytes
Expression Level
F
No. of associated pCREs
Gene Rank
Cyp2b10
Dner
Fgf1
Taf5
Igf1
Gata4
Nfkb2
Hes1
E
Serpina1a
Clec4g
Stab2
Dcn
Colec10
Clec4f
Ptprc
Itga4
Oasl1
Scd1
Spp1
Hepatocytes
LSECs
HSCs
Kupffer Cells
B Cells
Monocytes
Neurons
Cholangiocytes
G
All peaks
pCRES
Normalized Density
Distance to TSS
H
chr14:62830000-63320000
[range 0-50]
Genes
4931440J10Rik
Wdfy2
Defb42
Fdft1
Ints6
Gm26536
n-R5s47
Defb48 Defb47Defb43
Defb30
Ctsb
Neil2
Gata4
pCREs
Links
Spearman corr.

**Figure 2: Joint expression and chromatin accessibility profiling in primary liver nuclei**

(A) UMAP visualisation of WNN-integrated scRNAseq and scATACseq modalities of 19,664 liver nuclei. Nuclei are coloured by cell types.
(B) Comparison of UMIs/cell across different single-cell technologies. Red shading denotes all multi-omic technologies. Datasets are this study, SHARE-seq[13] (murine skin cells), sci-CAR[11] (murine kidney nuclei), SNARE-seq[12] (adult & neonatal mouse cerebral cortex nuclei), 10x 3' Expression[17] (murine liver nuclei) and sci-RNAseq3 [9](E16.5 mouse embryo nuclei).
(C) Comparison of unique fragments per cell across different single-cell technologies. Colouring as in (B). Datasets differing to (B) are 10x 3'scATAC[42] (murine liver nuclei) and sciATAC-seq[43] (murine liver nuclei).
(D) Normalised gene expression of representative marker genes per cell type.
(E) Aggregate ATAC-seq tracks at marker accessibility peaks per cell type.
(F) Genes ranked by number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene (±500kbp from TSS) in LSECs. Marked are transcription factors & regulators within the top 1% of genes with a critical role in LSECs.
(G) Significantly correlated pCREs are enriched for TSS proximity. Normalised density of all peaks versus pCREs within ±50kbp of nearest TSS.
(H) Aggregate scATAC-seq track of LSECs at the *Gata4* locus and 500kbp upstream region. Loops denote pCREs significantly correlated with *Gata4* and are coloured by Spearman correlation of respective pCRE–*Gata4* comparison

**Figure 3**

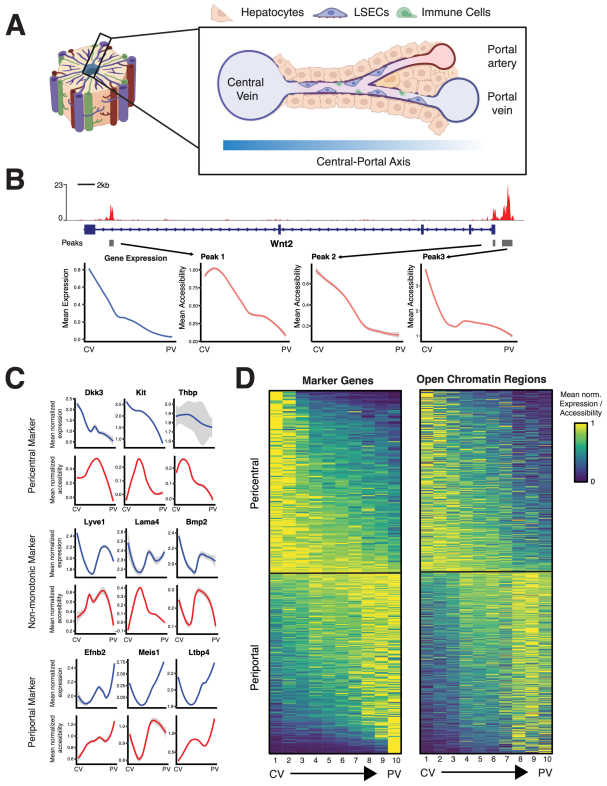



**Figure 3: Zonation profiles in LSECs across gene expression and chromatin accessibility**

(A) Schematic of a liver lobule. A liver lobule has a 'Central–Portal Axis' starting from the central vein to the portal vein and portal artery. The sinusoidal capillary channels are lined with LSECs.
(B) Changes along the Central–Portal Axis at the *Wnt2* locus. Top: Aggregate scATACseq profile (red) of LSECs at *Wnt2* locus. Grey bars denote identified peaks. Bottom: In blue, loess trend line of mean normalised *Wnt2* gene expression along the Central–Portal-Axis (central vein, CV; portal vein, PV; split into equal 10 bins). In red, loess trend line of mean normalised chromatin accessibility in peaks at the *Wnt2* locus along the CP-axis.
(C) Loess trend line of mean normalised expression (blue) and mean normalised accessibility along the Central–Portal axis for pericentral markers (top, increased toward the central vein, *Dkk3*, *Kit* and *Thbp*), non-monotonic markers (middle, increased between the veins, *Lyve1*, *Lama4* and *Bmp2*) and periportal markers (increased toward the portal vein, *Efnb2*, *Meis1* & *Ltbp4*)
(D) Left: Zonation profiles of 362 genes along the Central–Portal axis. Right: Zonation profiles of 846 open chromatin regions along the Central–Portal axis. All profiles are normalised by their maximum.

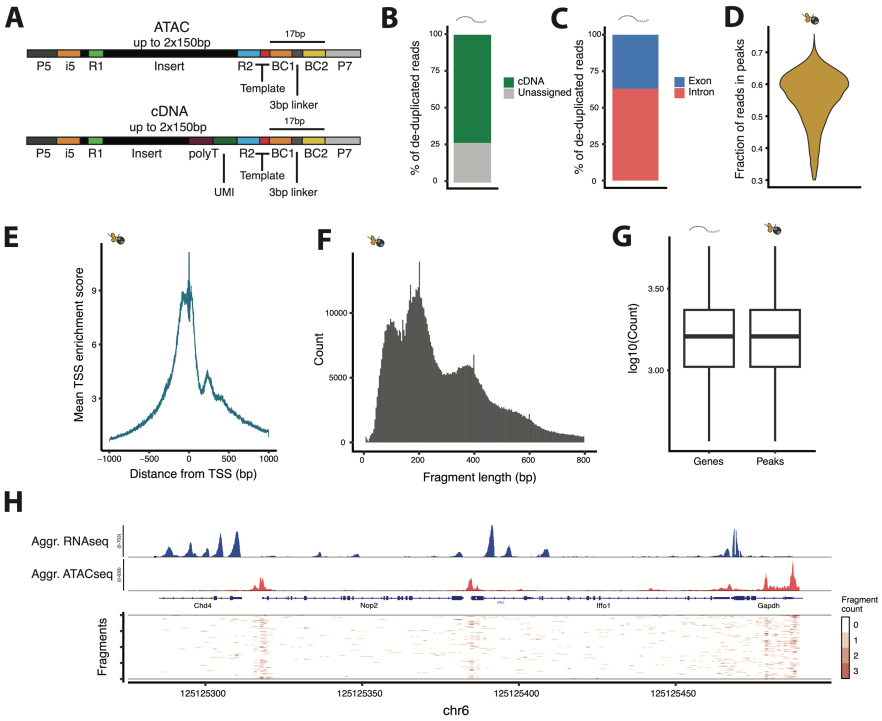
## Table 1

**Comparison of single-cell techniques**

| | Cost / Cell | Throughput | Multiomic? | Special equipment? | Std. sequencing? | Potential insert length? |
|---|---|---|---|---|---|---|
| This study | 5.6 ct | > 200.000 | Yes | No | Yes | > 200bp |
| SHARE-seq | 4.33 ct | > 200.000 | Yes | No | No | 100bp |
| 10x Multiome | 25.8 ct | 80.000 | Yes | Yes | No | 100bp |
| sci-RNA-seq3 | 1 ct | > 200.000 | No | No | Yes | > 200bp |

**Table 1: Comparison between different single-cell technologies**

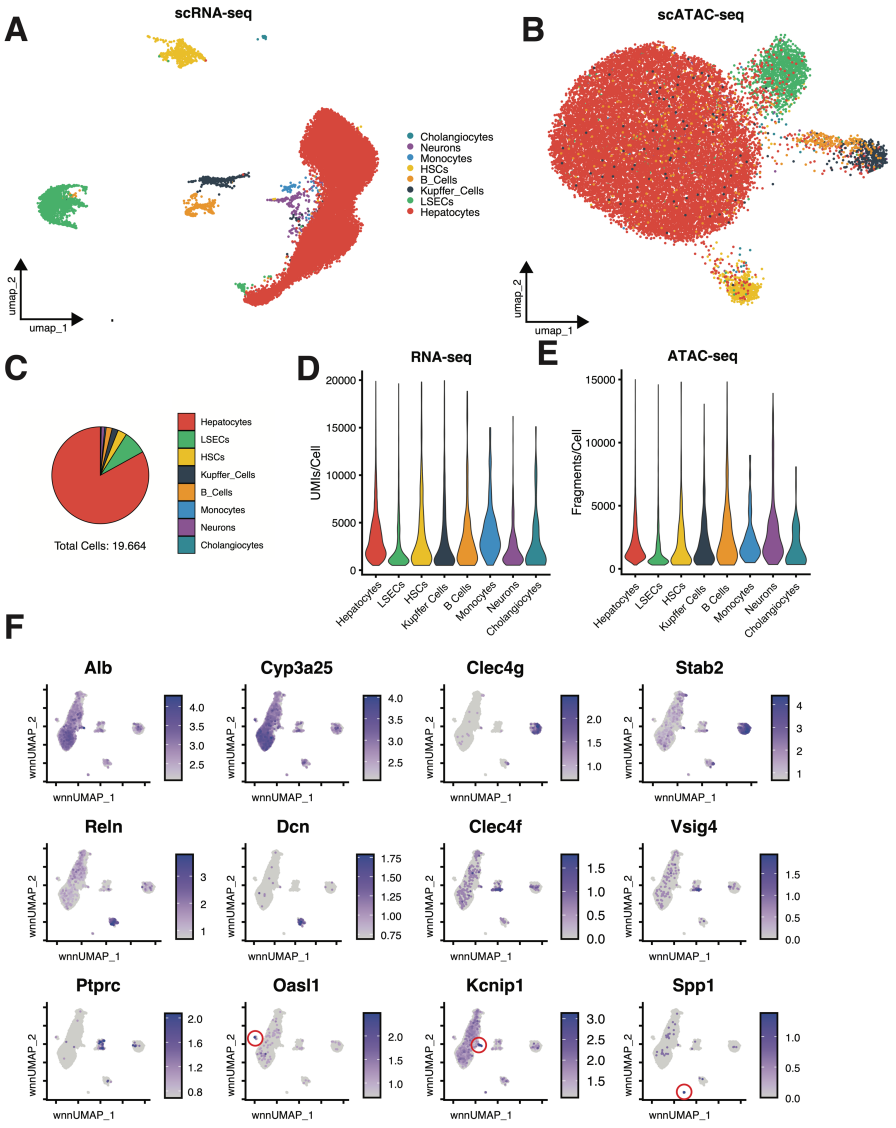
**Supplementary Figure 1**




**Supplementary Figure 1: Barcode structure and summary of quality control measures in liver nuclei**

(A) Structure of a scATAC-seq and scRNA-seq sequencing read. Created with Biorender.com
(B) Percentage of total scRNAseq sequencing reads containing cDNA fragments.
(C) Percentage of de-duplicated scRNAseq sequencing reads overlapping an exon or intron.
(D) Distribution of fraction of reads in peaks (FRiP) per cell in the scATAC-seq data (mean: 0.55).
(E) Mean TSS enrichment score per cell in relation to distance from nearest TSS in the scATACseq data.
(F) Histogram of fragment length in scATAC sequencing reads
(G) Expressed genes and accessible peaks per cell (mean expressed genes: 1,798; mean accessible peaks: 1,983)
(H) Top: Aggregate scRNA-seq (blue) and scATAC-seq (red) of all liver nuclei at *Nop2/Iffo2/Gapdh* locus. Bottom: Chromatin accessibility profiles of 100 individual cells.

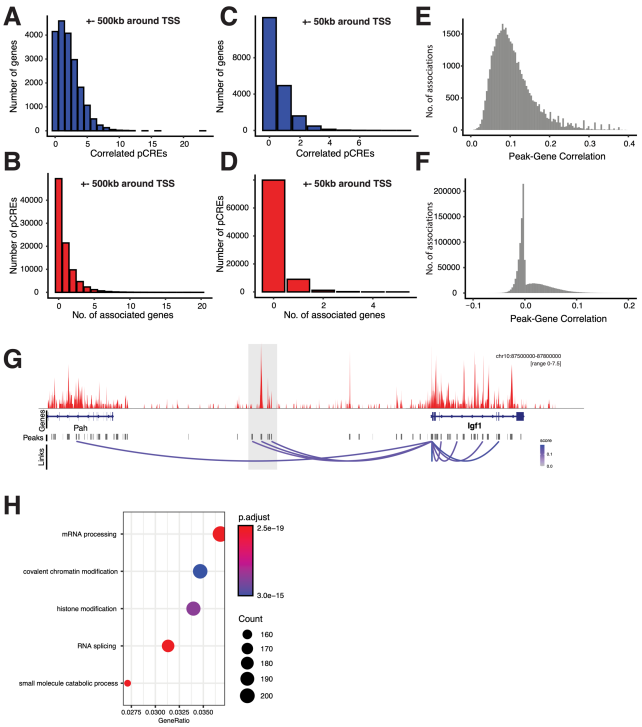
## Supplementary Figure 2



**Supplementary Figure 2: easySHAREseq robustly separates cell types**

(A) UMAP visualisation of merged and integrated scRNA-seq data. Nuclei are coloured according to their cell type.
(B) UMAP visualisation of merged and integrated scATAC-seq data. Nuclei are coloured according to their cell type.
(C) Fraction of cell types recovered relative to total cells
(D) Distribution of UMIs per cell split by cell type. Some cell types (e.g. LSECs) consistently yield less UMIs.
(E) Distribution of unique fragments per cell split by cell types. Some cell types (e.g. LSECs) consistently yield less fragments.
(F) WNN-UMAPs with cells coloured according to the mean expression strength of a given marker gene. Red circles indicate the position of the cell population showing elevated expression for this marker gene.

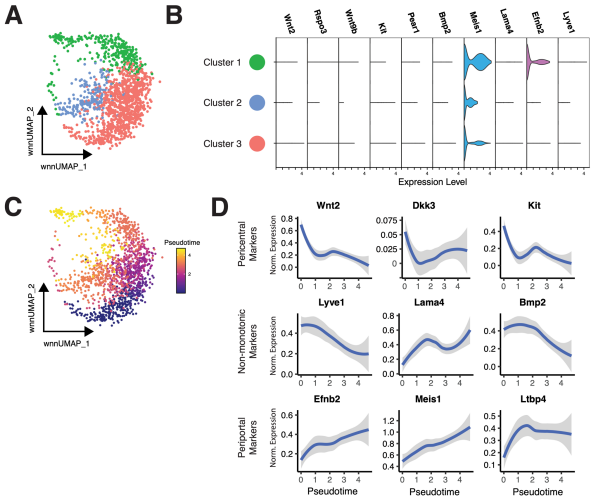
## Supplementary Figure 3




**Supplementary Figure 3: Summary of peak–gene correlations**

(A) Number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene, considering all peaks ±500kbp of the TSS
(B) Number of genes a given pCREs is significantly correlated with ($P < 0.05$, FDR = 0.1), considering all peaks ±500kbp of the TSS
(C) Number of significantly correlated pCREs ($P < 0.05$, FDR = 0.1) per gene, considering all peaks ±50kbp of the TSS
(D) Number of genes a given pCREs is significantly correlated with ($P < 0.05$, FDR = 0.1), considering all peaks ±50kbp of the TSS
(E) Histogram of Spearman correlations of all significant peak–gene correlations ($P < 0.05$)
(F) Histogram of Spearman correlations of all non-significant peak–gene correlations ($P > 0.05$)
(G) Aggregate scATAC-seq track of LSECs at the *Igf1* locus and its upstream region. Loops denote significantly correlated pCREs with *Igf1* and are coloured by their respective Spearman correlation. Shaded grey area denotes potentially LSEC-specific *cis*-regulatory element regulating *Igf1* expression.
(H) Gene Ontology enrichment analysis of genes whose associated pCREs are associated with five or more genes.

**Supplementary Figure 4**




**Supplementary Figure 4: Investigation of LSEC zonation**

(A) Subclustering LSECs reveals three distinct clusters.
(B) Comparison of marker gene expression across the three identified LSEC subclusters does not allow for fine-scale cell-type assignments.
(C) Subclustered LSECs coloured by pseudotime.
(D) Loess-Curve of marker gene expression of pericentral, non-monotonic and periportal marker genes along pseudotime.

## Methods

### Animal Model & Tissue preparation

#### *Mice*

All animal experimental procedures were carried out under the licence number EB 01-21M at Friedrich Miescher Laboratory of the Max Planck Society in Tübingen, Germany. The procedures were reviewed and approved by the Regierungspräsidium Tübingen, Germany. Liver was collected from both male and female wild-type C57BL/6 and PWD/PhJ mice aged between 9 to 11 weeks.

#### *Study design*

From each strain, we generated easySHARE-seq libraries for one male and one female mice from each strain (four total). For each individual, we sequenced two sub-libraries, resulting in 8 easySHAREseq libraries.

#### *Cell Culture*

For the species-mixing experiment, HEK Cells were cultured in media containing DMEM/F-12 with GlutaMAX™ Supplement, 10% FBS and 1% Penicillin-Streptomycin (PenStrep) at 37°C and 5% $CO_2$. Cells were harvested on the day of the experiment by simply pipetting them off the plate and were then spun down for 5 min at 250G.
For the second cell line, murine OP9-DL4 cells were cultured in alpha-MEM medium containing 5% FBS and 1% PenStrep. On the day of the experiment, the cells were harvested by aspirating the media and adding 4 ml of Trypsin, followed by an incubation at 37°C for 5 min. Then, 5ml of media was added and cells were spun down for 5 min at 250G.
After counting both cell lines using TrypanBlue and the Evos Countess II, equal cell numbers were mixed.

#### *Liver Nuclei*

The liver was extracted, rinsed in HBSS, cut into small pieces, frozen in liquid nitrogen and stored in the freezer at -80 °C for a maximum of two weeks. On the day of the experiment, 1 ml of ice cold Lysis Solution (0.1% Triton-X 100, 1mM DTT, 10mMM Tris-HCl pH8, 0.1mM EDTA, 3mM $Mg(Ac)_2$, 3mM $CaCl_2$ and 0.32M sucrose) was added to the tube. The cell suspension was transferred to a pre-cooled Douncer and dounced 10x using Pestle A (loose) and 15x using Pestle B (tight). The solution was added to a thick wall ultracentrifuge tube on ice and topped up with 4ml ice cold Lysis Solution. Then 9 ml of Sucrose solution (10mM Tris-HCl pH8.0, 3mM $Mg(Ac)_2$, 3mM DTT, 1.8M sucrose) was carefully pipetted to the bottom of the tube to create a sucrose cushion. Samples were spun in a pre-cooled ultracentrifuge with a SW-28 rotor at 24,400rpm for 1.5 hours at 4 °C. Afterwards, all supernatant was carefully aspirated so as not to dislodge the pellet at the bottom and 1 ml ice cold DEPC-treated water supplemented with 10µl SUPERase & 15µl Recombinant RNase Inhibitor was added. Without resuspending, the tube was kept on ice for 20 min. The pellet was then resuspended by pipetting ~15 times slowly up and down followed by a 40 µm cell straining step. Counting of the nuclei using DAPI and the Evos Countess II was immediately followed up by fixation.

**easySHARE-seq protocol**

***Preparing the barcoding oligonucleotides***

There are two barcoding rounds in easySHARE-seq with 192 unique barcodes distributed across two 96-well plates in each round (see **Suppl. Table 1** for a full list of oligonucleotide sequences). Each barcode (BC) is pre-annealed as a DNA duplex for improved stability. The first round of barcodes contains two single-stranded linker sequences at its ends as well as a 5' phosphate group to ligate the different barcodes together. The first single-stranded overhang links the barcode to a complementary overhang at the 5' end of the cDNA molecule or transposed DNA molecule, which originates either from the RT primer or the Tn5 adapter. The second overhang (3bp) is used to ligate it to the second round of barcodes (**Fig.1B**). Each duplex needs to be annealed prior to cellular barcoding, preferably on the day of the experiment. No blocking oligos are needed.

The Round1 BC plates contain 10µl of 4µM duplexes in each well and Round2 BC plates contain 10µl of 6µM barcode duplexes in each well, all in Annealing Buffer (10mM Tris pH8.0, 1mM EDTA, 30mM KCl). Pre-aliquoted barcoding plates can be stored at -20 °C for at least three months. On the day of the experiment, the oligo plates were thawed and annealed by heating plates to 95 °C for 2 min, followed by cooling down the plates to 20 °C at a rate of -2 °C per minute. Finally, the plates were spun down. Until the annealed barcoding plates are needed, they should be kept on ice or in the fridge.

This barcoding scheme is very flexible and currently supports a throughput of ~350,000 cells (assuming 96 indexing primers) per experiment, limited only by sequencing cost and availability of indexing primer. The barcodes were designed to have at least a Hamming distance of 2. See Supplementary Notes for further details on the barcoding system and flexibility.

**Tn5 *preparation***

Tn5 was expressed in-house as previously described [44]. Two differently loaded Tn5 are needed for easySHARE-seq, one for the tagmentation, loaded with an adapter for attaching the first barcodes (termed Tn5-B2S), and one for library preparation with a standard illumina sequencing adapter (termed Tn5-A-only). See Supplementary Table 1 for all sequences.

To assemble Tn5-B2S, two DNA duplexes were annealed: 20 µM Tn5-A oligo with 22 µM Tn5-reverse and 20 µM Tn5-B2S with 22 µM Tn5-reverse, all in 50 mM NaCl and 10mM Tris pH8.0. Oligos were annealed by heating the solution to 95 °C for 30 s and cooling it down to 20 °C at a rate of 2 °C/min. An equal volume of duplexes was pooled and then 200 µl of unassembled Tn5 was mixed with 16.5 µl of duplex mix. The Tn5 was then incubated at 37 °C for 1 hour, followed by 4 °C overnight. The Tn5 can then be stored at -20 °C. In our hands, Tn5 did not show a decrease in activity after 10 months of storage.

To assemble Tn5-A-only, 10 µM of Tn5-A and 10.5 µM Tn5-reverse was annealed using the same conditions as described above. Again, 200 µl of unassembled Tn5 was mixed with 16.5 µl of Tn5-A duplex and incubated at 37 °C for 1 hour, followed by 4 °C overnight. The Tn5 can then be stored for later and repeated use for more than 10 months at -20 °C.

We observed an increase in all Tn5 activity during the first months of storage, possibly due to continued transposome assembly in storage.

***Fixation***

One million liver nuclei ("cells" for short) were added to ice-cold PBS for 4 ml total. After mixing, 87 µl 16% formaldehyde solution (0.35%; for liver nuclei) or 25 µl 16% formaldehyde solution

(0.1%; for HEK and OP9 cells) was added and the suspension was mixed by pipetting up and down exactly 3 times with a P1000 pipette set to 700 µl. The suspension was incubated at room temperature for 10 min. Fixation was stopped by adding ice-cold Stop-Mix (224 µl 2.5M glycine, 200 µl 1M Tris-HCl pH8.0, 53 µl 7.5% BSA in PBS). The suspension was mixed exactly 3 times with a P1000 pipette set to 850 µl and incubated on ice for 3 min followed by a centrifugation at 500G for 5 min at 4°C. Supernatant was removed and the pellet was resuspended in 1 ml Nuclei Isolation Buffer (NIB; 10mM Tris pH8.0, 10mM NaCl, 2mM $MgCl_2$, 0.1% NP-40) and kept on ice for 3 min followed by straining the suspension with a 40 µm cell strainer. It was then spun down at 500G for 3 min at 4°C and re-suspended in ~100-200µl PBSi (1x PBS + 0.4 U/µl Recombinant RNaseInhibitor, 0.04% BSA, 0.2 U/µl SUPERase, freshly added), depending on the amount of input cells. Cells were then counted using DAPI and the Countess II and concentration was adjusted to 2M cells/ml using PBSi.

***Tagmentation***

In a typical easySHARE-seq experiment for this study, 8 tagmentation reactions with 10,000 cells each followed by 3 RT reactions were performed. This results in sequencing libraries for around 30,000 cells. To increase throughput, simply increase the amount of tagmentation and RT reactions accordingly. No adjustment is needed to the barcoding. Each tube and PCR strip until the step of Reverse Crosslinking was coated before use by rinsing it with PBS+0.5% BSA.

For each tagmentation reaction, 5 µl of 5X TAPS-Buffer, 0.25µl 10% Tween, 0.25µl 1% Digitonin, 3 µl PBS, 1 µl Recombinant RNaseInhibitor and 9µl of H2O was mixed. TAPS Buffer was made by first making a 1M TAPS stock solution in $H_2O$, followed by adjustment of the pH to 8.5 by titrating 10M NaOH. Then, 4.25ml $H_2O$, 500µl 1M TAPS pH8.5, 250µl 1M $MgCl_2$ and 5ml N-N-Di-Methyl-Formamide (DMF) was mixed on ice and in order. When adding DMF, the buffer heats up so it is important to be kept on ice. The resulting 5X TAPS-Buffer can then be stored at 4°C for short term use (1-2 months) or for long-term storage at -20°C (> 6 months).

Then, 5 µl of cell suspension at 2M cells/ml in PBSi was added to the tagmentation mix for each reaction, mixed thoroughly and finally 1.5µl of Tn5-B2S was added. The reaction was incubated on a shaker at 37°C for 30 min at 850 rpm. Afterwards, all reactions were pooled on ice into a pre-cooled 15ml tube. The reaction wells were washed with ~30 µl PBSi which was then added to the pooled suspension in order to maximize cell recovery. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the cells were washed with 200µl NIB followed by another centrifugation at 500G for 3 min at 4°C.

We only observed cell pellets when centrifuging after fixation and only when using cell lines as input material. Therefore, when aspirating supernatant at any step it is preferable to leave around 20-30µl liquid in the tube. Additionally, it is recommended to pipette gently at any step as to not damage and fracture the cells.

***Reverse Transcription***

As stated above, three tagmentation reactions were combined into one RT reaction. When increasing cells to more than 30,000 per RT reaction, we observed a steep drop in reaction efficiency.

The Master Mix for one RT reaction contained 3µl 100µM RT-primer, 2µl 10mM dNTPs, 6µl 5X MaximaH RT Buffer, 4.5µl 50% PEG6000, 1.5 µl $H_2O$, 1.5µl SUPERase and 1.66µl MaximaH RT. The RT primer contains a polyT tail, a 10bp UMI sequence, a biotin molecule and an adapter sequence used for ligating onto the first round of barcoding oligos.

The cell suspension was resuspended in 10µl NIB per RT reaction and added to the Master Mix for a total of 30µl. As PEG is present, it is necessary to pipette ~30 times up and down to ensure proper mixing. The RT reaction was performed in a PCR cycler with the following protocol: 52°C for 12 min; then 2 cycles of 8°C for 12s, 15°C for 45s, 20°C for 45s, 30°C for 30s, 42°C for 2min and 50°C for 3 min. Finally, the reaction was incubated at 52°C for 5 more minutes. All reactions were then pooled on ice into a pre-cooled and coated 15ml tube and the reaction wells were washed with ~40µl NIB, which was then added to the pooled cell suspension in order to maximise cell recovery. The suspension was then spun down at 500G for 3 min at 4°C. Supernatant was aspirated and the cells were washed in 150µl NIB and spun down again at 500G for 3min at 4°C. This washing step was repeated once more, followed by resuspension of the cells in 2ml Ligation Mix (400µl 10x T4-Buffer, 40µl 10% Tween-20, 1460µl Annealing Buffer and 100µl T4 DNA Ligase, added last).

***Single-cell barcoding***

Using a P20 pipette, 10µl of cell suspension in the ligation mix was added to each well of the two annealed Round1 BC plates, taking care as to not touch the liquid at the bottom of each well. The plates were then sealed, shaken gently by hand and quickly spun down (~ 8s) followed by an incubation on a shaker at 25°C for 30 min at 350 rpm. After 30 min, the cells from each well were pooled into a coated PCR strip using a P200 multichannel pipette set to 30µl. In order to pool, each row was pipetted up and down three times before adding the liquid to the PCR strip. After 8 columns were pooled into the strip, the suspension was transferred into a coated 5ml tube on ice. This process was repeated until both plates were pooled, taking care to aspirate most liquid from the plates. The cell suspension was then spun down for 3min at 500G at 4°C. Supernatant was aspirated and the cells were resuspended thoroughly in 2 ml new Ligation Mix. Now, 10µl of cell suspension was added into each well of the annealed Round2 barcoding plates using a P20 pipette, taking care as to not touch the liquid within each well. The plates were sealed, shaken gently by hand and spun down quickly followed by incubating them on a shaker at 25°C for 45 min at 350 rpm. The cells were then pooled again using the above described procedure into a new coated 15ml Tube. The cells were spun down at 500G for 3 min at 4°C. Supernatant was aspirated, the cells were washed with 150µl NIB and spun down again. Finally, the cells were resuspended in ~60µl NIB and counted. For counting, 5µl of cells were mixed with 5µl of NIB and 1x DAPI and counted on the Evos Countess II, taking the dilution into account. Sub-libraries of 3,500 cells were made and the volume was adjusted to 25µl by addition of NIB.
Using 3,500 cells results in a doublet rate of ~6.3%. The recovery rate of cells after sequencing depends on the input material (and QC thresholds), with cell lines recovering around 80% of input cells (~2,800-3,000 cells) and liver nuclei around 70% (~2,300-2,500 cells).

***Reverse-Crosslinking***

To each sub-library of 3,500 cells, 30µl 2x Reverse Crosslinking (RC) Buffer (0.4% SDS, 100mM NaCl, 100mM Tris pH8.0) as well as 5µl ProteinaseK was added. The sub-libraries were mixed and incubated on a shaker at 62°C for one hour at 800 rpm. Afterwards, they were transferred to a PCR cycler into a deep well module set to 62°C (lid to 80°C) for an additional hour. Afterwards, each sub-library was incubated at 80°C for 10 min and finally 5µl of 10% Tween-20 to quench the SDS and 35µl of NIB was added for a total volume of 100µl.
The lysates can be stored at this point at -20°C for at least two days, which greatly simplifies handling many sub-libraries at once. Longer storage has not been extensively tested.

***Streptavidin Pull-Down***

Each transcript contains a biotin molecule as the RT primers are biotinylated which is used to separate the scATAC-seq libraries from the scRNA-seq libraries. For each sublibrary, 50µl M280 Streptavidin beads were washed three times with 100µl B&W Buffer (5mM Tris pH8.0, 1M NaCl, 0.5mM EDTA) supplemented with 0.05% Tween-20, using a magnetic stand. Afterwards, the beads were resuspended in 100µl 2x B&W Buffer and added to the sublibrary, which were then shaken at 25°C for one hour at 900 rpm. Now all cDNA molecules are attached to the beads whereas transposed molecules are within the supernatant. The lysate was put on a magnetic stand to separate supernatant and beads.

It likely is possible to reduce the number of M280 beads in this step, significantly reducing overall costs. However, this has not been extensively tested.

***scATAC-seq library preparation***

The supernatant from each sub-library was cleaned up with a Qiagen MinElute Kit and eluted twice into 30µl 10mM Tris pH8.0 total. PCR Mix containing 10µl 5X Q5 Reaction Buffer, 1µl 10mM dNTPs, 2µl 10µM i7-TruSeq-long primer, 2µl 10µM Nextera N5XX Indexing primer, 4.5µl H2O and 0.5µl Q5 Polymerase was added (All Oligo sequences in **Suppl. Table 1**). Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The fragments were amplified with the following protocol: 72°C for 6 min, 98°C for 1 min, then cycles of 98°C for 10s, 66°C for 20s and 72°C for 45s followed by a final incubation at 72°C for 2 min. The number of PCR cycles strongly depends on input material (Liver: 17 PCR cycles, Cell Lines: 15 PCR cycles). The reactions were then cleaned up with custom size selection beads with 0.55X as upper cutoff and 1.4X as lower cutoff and eluted into 25µl 10mM Tris pH8.0. Libraries were quantified using the Qubit HS dsDNA Quantification Kit and run on the Agilent 2100 bioanalyzer with a High Sensitivity DNA Kit.

***cDNA library preparation***

The beads containing the cDNA molecules were washed three times with 200µl B&W Buffer supplemented with 0.05% Tween-20 before being resuspended in 100µl 10mM Tris ph8.0 and transferred into a new PCR strip. The strip was put on a magnet and the supernatant was aspirated. The beads were then resuspended in 50µl Template Switch Reaction Mix: 10µl 5X MaximaH RT Buffer, 2µl 100µM TS-oligo, 5µl 10mM dNTPs, 3µl Enzymatics RNaseIn, 15µl 50% PEG6000, 14µl H2O and 1.25µl MaximaH RT. The sample was mixed well and incubated at 25°C for 30 min followed by an incubation at 42°C for 90 min. The beads were then washed with 100µl 10mM Tris while the strip was on a magnet and resuspended in 60µl $H_2O$. To each well, 40µl PCR Mix was added containing 20µl 5X Q5 Reaction Buffer, 4µl 10µM i7-Tru-Seq-long primer, 4µl 10µM Nextera N5XX Indexing primer, 2µl 10mM dNTPs, 9µl H2O and 2µl Q5 Polymerase. The resulting mix can be split into two 50µl PCR reactions or run in one 100µl reaction. The PCR involved initial incubation at 98°C for 1 min followed by PCR cycles of 98°C for 10s, 66°C for 20s and 72°C for 3 min with a final incubation at 72°C for 5 min. Importantly, in order to distinguish the samples, each sub-library needs to be indexed with a different N5XX Indexing primer. The number of PCR cycles strongly depends on input material (Liver: 15 cycles, Cell lines: 13 cycles).

The PCR reactions were cleaned up with custom size selection beads using 0.7X as a lower cutoff (70µl) and eluted into 25µl 10mM Tris pH8.0. The cDNA libraries were quantified using the Qubit HS dsDNA Quantification Kit.

***Species-Mixing Experiments***

RNA-seq reads were aligned to a composite hg38-mm10 genome using STAR[46]. The resulting bamfile was then filtered for uniquely mapping reads and reads mapping to chrM, chrY or unmapped scaffolds or containing unplaced barcodes were removed. Finally, the reads were deduplicated using UMItools[45]. ATAC-seq reads were also aligned to a composite genome using bwa[47]. Duplicates were removed with Picard tools and reads mapping to chrM, chrY or unmapped scaffolds were filtered out. Additionally, reads that were improperly paired or had an alignment quality < 30 were also removed.

The reads were then split depending on which genome they mapped to and reads per barcode were counted. Barcodes needed to be associated with at least 700 fragments and 500 UMIs in order to be considered a cell for the analysis. A barcode was considered a doublet when either the proportion of UMIs or fragments assigned to a species was less than 75%. This cutoff was chosen to mitigate possible mapping bias within the data.

***easySHARE-RNA-seq processing and read alignment***

We only used Read 1 for all our RNA-seq analyses as sequencing quality tends to drop after a polyT tail is sequenced in R2. Each sample was mapped to mm10 using the twopass mode in STAR[46] with the parameters --outFilterMultimapNmax 20 --outFilterMismatchNmax 15. We then processed the bamfiles further by moving the UMI and barcode from the read name to a bam flag, filtering out multimapping reads and reads without a definitive barcode. To determine if a read overlapped a transcript, we used featureCounts from the subread package[48]. UMI-Tools was used to collapse the UMIs of aligned reads, allowing for one mismatch and de-duplication of the reads. Finally, (single-cell) count matrices were created also using UMI-Tools.

***easySHARE-ATAC-seq pre-processing and read alignment***

Fastq files were demultiplexed using a custom C-script, allowing one mismatch within each barcode segment. The paired reads were trimmed using cutadapt[49] and the resulting reads were mapped to the mm10 genome using bwa mem[47]. Reads with alignment quality < Q30, unmapped, undetermined barcode, or mapped to mtDNA were discarded. Duplicates were removed using Picard tools. Open chromatin regions were called by subsampling the bamfiles from all samples to a common depth, merging them into a pooled bamfile and using the peak caller MACS2[50] with the parameters -nomodel -keep-dup -min-length 100. The count matrices as well as the FRiP score was generated using featureCounts from the Subread package[48] together with the tissue-specific peak set.

***Filtering, Integration & Dimensional reduction of scRNAseq data***

The count matrices were loaded into Seurat[51] and cells were then filtered for >200 detected genes, >500 UMIs and < 20.000 UMIs. The sub-libraries coming from the same experiment were then merged together and normalised. Merged experiments from the same species (one from male mouse, one from female mouse) were then integrated by first using SCTransform[52] to normalise the data, then finding common features between the two experiments using FindIntegrationAnchors() and finally integrated using IntegrateData(). Lastly, the integrated datasets from C57BL/6 and PWD/PhJ were again integrated using IntegrateData(). To visualise the data, we projected the cells into 2D space by UMAP using the first 30 principal components and identified clusters using FindClusters().

***Filtering, Integration & Dimensional reduction of scATACseq data***

Fragments per cell were counted using sinto and the resulting fragment file was loaded into Signac[53] alongside the count matrices and the peakset. We calculated basic QC statistics using base Signac and cells were then filtered for a FRiP score of at least 0.3, > 300 fragments, < 15.000 fragments, a TSS enrichment > 2 and a nucleosome signal < 4. Again, sublibraries coming from the same experiment were merged. We then integrated all four experiments (C57BL/6 & PWD/PhJ, one male & one female mouse each) by finding common features across datasets using FindIntegrationAnchors() using PCs 2:30 and then integrating the data using IntegrateEmbeddings(). To visualise the data, we projected the cells into 2D space by UMAP.

***Weighted-Nearest-Neigbor (WNN) Analysis & Cell type identification***

In order to use data from both modalities simultaneously, we created a multimodal Seurat object and used WNN[16] clustering to visualise and leverage both modalities for downstream analysis. Afterwards, we assigned cell cycle scores and excluded clusters consisting of nuclei solely in the G2M-phase (2 clusters, 121 nuclei total). Cell types were assigned via expression of previously known marker genes, which allows subsetting the data into cell types.

***Calculating Peak–Gene Associations***

Peak–gene associations were calculated following the framework described by Ma et al[13]. In short, Spearman correlation was calculated for every peak–gene pair within a +-500kb window around the TSS of the expressed gene. To obtain a background estimation, we used chromVAR[54] (*getBackgroundPeaks*()) to generate 100 background peaks matched in GC bias and chromatin accessibility but randomly distributed throughout the genome. We calculated the Spearman correlation between every background-gene comparison, resulting in a null distribution with known population mean and standard deviation. We then calculated the z-score for the peak–gene pair in question ((correlation - population mean)/ standard deviation) and used a one-sided z-test to determine the p-value. This functionality is also implemented in Signac under the function *LinkPeaks*(). Increasing the number of background peaks to 200, 350 or 500 for each peak–gene pair does not impact the results (*data not shown*).

***Analysis of LSEC zonation markers***

To analyse gene expression and chromatin accessibility along LSEC zonation, we subsetted our data for LSECs only, extracted expression values and wnnUMAP coordinates and binned the data along the wnnUMAP_2 axis into 10 equal sized bins. We then calculated the mean expression/accessibility for each gene/peak in each bin, excluding cells that contained a zero count. To identify novel marker genes, we excluded genes with low expression and calculated the moving average (for three bins) across the bins. We then required the moving average to continuously decrease (for pericentral marker genes) or increase ( for periportal marker genes), allowing two exceptions. Lastly, we divided the means for each gene by their maximum to normalise the values. Identification of *cis*-regulatory elements displaying zonation effects had equal requirements.

Imputation of pseudotime was performed in Monocle3[55] with standard parameters. Gene expression was smoothed over both bins and pseudotime (separately) with local polynomial regression fitting (loess).

***Gene Ontology Analysis***

Gene Ontology Analysis was done using the R package clusterProfiler[56] with standard parameters.

***Data Availability***

All data can be accessed using the accession number GSE256434. All code used in data analysis is available at https://github.com/vosoltys/easySHARE_seq.git.

***Acknowledgements***

We thank members of the Chan and Jones lab for helpful discussions and critical reading of the manuscript. We are very grateful to Arnar Breevoort and Alex Pollen for sharing tissue preparation protocols and a very helpful research visit. We thank Sinja Mattes and all animal care takers at the Friedrich Miescher Laboratory for their work. We also thank the Genome Center in the Max Planck Institute for Biology for providing support.

***Author Contributions***

V.S. and M.P. developed the barcoding framework for easySHAREseq. V.S. developed the rest of the protocol and performed all experiments. V.S. did all computational analyses and wrote the manuscript. M.P., D.S., M.K. and Y.F.C. helped with experimental or computational support. All authors reviewed the manuscript.

***Declaration of Interest***

The authors declare no competing interests.

## References

1. Anderson, E., Devenney, P. S., Hill, R. E. & Lettice, L. A. Mapping the Shh long-range regulatory domain. *Development* **141**, 3934–3943 (2014).
2. Nord, A. S. *et al.* Rapid and Pervasive Changes in Genome-wide Enhancer Usage during Mammalian Development. *Cell* **155**, 1521–1531 (2013).
3. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**, 613–626 (2012).
4. Buenrostro, J. D. *et al.* Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation. *Cell* **173**, 1535-1548.e16 (2018).
5. Zhang, C., Macchi, F., Magnani, E. & Sadler, K. C. Chromatin states shaped by an epigenetic code confer regenerative potential to the mouse liver. *Nat Commun* **12**, 4110 (2021).
6. Lara-Astiaso, D. *et al.* Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
7. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).
8. Kashima, Y. *et al.* Single-cell sequencing techniques from individual to multiomics analyses. *Exp Mol Med* **52**, 1419–1427 (2020).
9. Martin, B. K. *et al.* Optimized single-nucleus transcriptional profiling by combinatorial indexing. *Nat Protoc* **18**, 188–207 (2023).
10. Rosenberg, A. B. *et al.* Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science* **360**, 176–182 (2018).
11. Cao, J. *et al.* Joint profiling of chromatin accessibility and gene expression in thousands of single cells. *Science* **361**, 1380–1385 (2018).
12. Chen, S., Lake, B. B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat Biotechnol* **37**, 1452–1457 (2019).
13. Ma, S. *et al.* Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell* **183**, 1103-1116.e20 (2020).
14. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biology* **21**, 31 (2020).
15. Aizarani, N. *et al.* A Human Liver Cell Atlas reveals Heterogeneity and Epithelial Progenitors. *Nature* **572**, 199–204 (2019).
16. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).
17. Su, Q. *et al.* Single-cell RNA transcriptome landscape of hepatocytes and non-parenchymal cells in healthy and NAFLD mouse liver. *iScience* **24**, 103233 (2021).
18. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).
19. Chen, X. *et al.* Structural insights into preinitiation complex assembly on core promoters. *Science* **372**, eaba8490 (2021).
20. Winkler, M. *et al.* Endothelial GATA4 controls liver fibrosis and regeneration by preventing a pathogenic switch in angiocrine signaling. *J Hepatol* **74**, 380–393 (2021).
21. Géraud, C. *et al.* GATA4-dependent organ-specific endothelial differentiation controls liver development and embryonic hematopoiesis. *J Clin Invest* **127**, 1099–1114.
22. Lara-Diaz, V. *et al.* IGF-1 modulates gene expression of proteins involved in inflammation, cytoskeleton, and liver architecture. *J Physiol Biochem* **73**, 245–258 (2017).

23. Baratta, J. L. *et al.* Cellular Organization of Normal Mouse Liver: A Histological, Quantitative Immunocytochemical, and Fine Structural Analysis. *Histochem Cell Biol* **131**, 713–726 (2009).
24. Jungermann, K. & Kietzmann, T. Zonation of parenchymal and nonparenchymal metabolism in liver. *Annu Rev Nutr* **16**, 179–203 (1996).
25. Braeuning, A. *et al.* Differential gene expression in periportal and perivenous mouse hepatocytes. *The FEBS Journal* **273**, 5051–5061 (2006).
26. Planas-Paz, L. *et al.* The RSPO–LGR4/5–ZNRF3/RNF43 module controls liver zonation and size. *Nat Cell Biol* **18**, 467–479 (2016).
27. Wang, B., Zhao, L., Fish, M., Logan, C. Y. & Nusse, R. Self-renewing diploid Axin2+ cells fuel homeostatic renewal of the liver. *Nature* **524**, 180–185 (2015).
28. Halpern, K. B. *et al.* Paired-cell sequencing enables spatial gene expression mapping of liver endothelial cells. *Nat Biotechnol* **36**, 962–970 (2018).
29. Knolle, P. A. & Wohlleber, D. Immunological functions of liver sinusoidal endothelial cells. *Cell Mol Immunol* **13**, 347–353 (2016).
30. Smedsrød, B. Clearance function of scavenger endothelial cells. *Comparative Hepatology* **3**, S22 (2004).
31. Rafii, S., Butler, J. M. & Ding, B.-S. Angiocrine functions of organ-specific endothelial cells. *Nature* **529**, 316–325 (2016).
32. Theilmann, A. L. *et al.* Endothelial BMPR2 Loss Drives a Proliferative Response to BMP (Bone Morphogenetic Protein) 9 via Prolonged Canonical Signaling. *Arteriosclerosis, Thrombosis, and Vascular Biology* **40**, 2605–2618 (2020).
33. Russell, K. S., Stern, D. F., Polverini, P. J. & Bender, J. R. Neuregulin activation of ErbB receptors in vascular endothelium leads to angiogenesis. *American Journal of Physiology-Heart and Circulatory Physiology* **277**, H2205–H2211 (1999).
34. Vihanto, M. M. *et al.* Hypoxia up-regulates expression of Eph receptors and ephrins in mouse skin. *FASEB J* **19**, 1689–1691 (2005).
35. Shen, Z. *et al.* Delta-Like Ligand 4 Modulates Liver Damage by Down-Regulating Chemokine Expression. *Am J Pathol* **186**, 1874–1889 (2016).
36. Zellmer, S. *et al.* Transcription factors ETF, E2F, and SP-1 are involved in cytokine-independent proliferation of murine hepatocytes. *Hepatology* **52**, 2127–2136 (2010).
37. Dong, X. C. *et al.* Inactivation of Hepatic Foxo1 by Insulin Signaling Is Required for Adaptive Nutrient Homeostasis and Endocrine Growth Regulation. *Cell Metabolism* **8**, 65–76 (2008).
38. Wang, X., Yu, Y., Xie, H.-B., Shen, T. & Zhu, Q.-X. Complement regulatory protein CD59a plays a protective role in immune liver injury of trichloroethylene-sensitized BALB/c mice. *Ecotoxicology and Environmental Safety* **172**, 105–113 (2019).
39. Ren, H. *et al.* Sirtuin 2 Prevents Liver Steatosis and Metabolic Disorders by Deacetylation of Hepatocyte Nuclear Factor 4α. *Hepatology* **74**, 723 (2021).
40. Miyao, M. *et al.* Pivotal role of liver sinusoidal endothelial cells in NAFLD/NASH progression. *Laboratory Investigation* **95**, 1130–1144 (2015).
41. Su, T. *et al.* Single-Cell Transcriptomics Reveals Zone-Specific Alterations of Liver Sinusoidal Endothelial Cells in Cirrhosis. *Cellular and Molecular Gastroenterology and Hepatology* **11**, 1139–1161 (2021).
42. Nikopoulou, C. *et al.* Spatial and single-cell profiling of the metabolome, transcriptome and epigenome of the aging mouse liver. *Nat Aging* **3**, 1430–1445 (2023).
43. Cusanovich, D. A. *et al.* A Single-Cell Atlas of In Vivo Mammalian Chromatin Accessibility. *Cell* **174**, 1309-1324.e18 (2018).

44. Picelli, S. *et al.* Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. *Genome Res.* **24**, 2033–2040 (2014).
45. Smith, T., Heger, A. & Sudbery, I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res* **27**, 491–499 (2017).
46. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
50. Zhang, Y. *et al.* Model-based Analysis of ChIP-Seq (MACS). *Genome Biology* **9**, R137 (2008).
51. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).
52. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**, 296 (2019).
53. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat Methods* **18**, 1333–1341 (2021).
54. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat Methods* **14**, 975–978 (2017).
55. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).
56. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *OMICS: A Journal of Integrative Biology* **16**, 284–287 (2012).

# Appendix 2

## Rapid genotype imputation from sequence with reference panels

Robert W Davies [1], Marek Kucka [2], Dingwen Su [2], Sinan Shi [1], Maeve Flanagan [3], Christopher M Cunniff [3], Yingguang Frank Chan [#2], Simon Myers [#1]

1 Department of Statistics, University of Oxford, Oxford, United Kingdom

2 Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany 3 Department of Pediatrics, Weill Cornell Medical College, New York, USA

# These authors contributed equally to this work.

Supplementary material available at https://doi.org/10.1038/s41588-021-00877-0

TECHNICAL REPORT
https://doi.org/10.1038/s41588-021-00877-0

nature genetics

Check for updates

# Rapid genotype imputation from sequence with reference panels

**Robert W. Davies[1]✉, Marek Kucka[2], Dingwen Su[2], Sinan Shi[1], Maeve Flanagan[3], Christopher M. Cunniff[3], Yingguang Frank Chan[2,5] and Simon Myers[1,4,5]**

**Inexpensive genotyping methods are essential to modern genomics. Here we present QUILT, which performs diploid genotype imputation using low-coverage whole-genome sequence data. QUILT employs Gibbs sampling to partition reads into maternal and paternal sets, facilitating rapid haploid imputation using large reference panels. We show this partitioning to be accurate over many megabases, enabling highly accurate imputation close to theoretical limits and outperforming existing methods. Moreover, QUILT can impute accurately using diverse technologies, including long reads from Oxford Nanopore Technologies, and a new form of low-cost barcoded Illumina sequencing called haplotagging, with the latter showing improved accuracy at low coverages. Relative to DNA genotyping microarrays, QUILT offers improved accuracy at reduced cost, particularly for diverse populations that are traditionally underserved in modern genomic analyses, with accuracy nearly doubling at rare SNPs. Finally, QUILT can accurately impute (four-digit) human leukocyte antigen types, the first such method from low-coverage sequence data.**

Large genome-wide association studies (GWAS) are essential to modern human genomics. They pinpoint individual genes that contribute to specific phenotypes, allow for accurate measuring of disease heritability, reveal genetic relationships between phenotypes and allow for dissection of the contribution of tissues to specific phenotypes[1]. In addition, large GWAS are the basis for generating accurate polygenic risk scores[2], which are essential to realizing the promise of precision medicine[3].

Over the past decade, most GWAS have first genotyped half a million or more SNPs using a genotyping microarray and then imputed additional untyped SNPs using haplotype reference panels[4,5]. The algorithms that perform the phasing and imputation are generally derived from the Li and Stephens model[6], which models each sample individual as a mosaic of reference haplotypes. In such models, imputation accuracy increases with reference panel size due to longer and hence more recent matches between sample and reference haplotypes, leading to increasingly large reference panels, for example, the haplotype reference consortium (HRC)[7]. To handle these large panels, sophisticated phasing and imputation methods have been developed to work specifically from genotyping microarray input, resulting in fast run times and very high accuracy[5,8–11].

Recently, low-coverage whole-genome sequencing (lcWGS) has emerged as an alternative to genotyping microarrays for obtaining genotyping information for imputation[12–14]. This approach has become increasingly attractive since both library[15,16] and sequencing costs have decreased; especially for non-model organisms, it avoids expensive array design or low-throughput costs[17]. Methods to impute from lcWGS include some derived from approaches designed primarily for array-based imputation[18], with others designed specifically for lcWGS[17,19,20] or designed principally around nonhuman imputation[21–23].

However, none of these methods are designed specifically for lcWGS using large haplotype reference panels. Data from sequencing reads from lcWGS have different properties than probes from genotyping microarrays and warrant different statistical models for phasing and imputation. First, probes from genotyping microarrays are short, being tens of base pairs (bp) long; as such, information from different probes, covering different SNPs, is nearly always independent. By contrast, sequencing reads are typically 100–250 bp long, may be paired and with long-read sequencing, may be many thousands of bp long. As such, information at nearby SNPs may not be independent by coming from the same read(s). This issue is particularly acute when reads are expected to span more SNPs on average, for example, in highly polymorphic regions like the major histocompatibility complex, in populations of species with high genetic diversity, and with long-read sequencing technologies, such as Oxford Nanopore Technologies (ONT) and Pacific Biosciences. Second, information from genotyping microarrays is summed across paternal and maternal haplotypes, while sequencing reads come from either the paternal or maternal haplotype. If the maternal or paternal origin of each sequencing read can be determined, imputation becomes much simpler because maternal and paternal reads can be separated and haploid imputation performed, which has linear or better computational complexity in reference panel size. This reduction of complexity can greatly speed up computation and ensures efficient processing of many thousands of samples in large studies.

In this study, we present QUILT, a method for rapid genotype imputation and phasing from lcWGS using a large haplotype reference panel. QUILT uses efficient computational storage of reference haplotypes and an iterative two-step procedure with Gibbs sampling to efficiently impute samples from lcWGS with linear computational complexity in the number of samples, SNPs and reference haplotypes. We evaluate QUILT's performance versus genotyping microarrays on three exemplar data types, representing the diversity of sequencing approaches currently available: Illumina short-read sequencing data; ONT, one example of long-read sequencing data; and haplotagging data, a low-cost, scalable form of barcoded

[1]Department of Statistics, University of Oxford, Oxford, UK. [2]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen, Germany. [3]Department of Pediatrics, Weill Cornell Medical College, New York, NY, USA. [4]The Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK. [5]These authors contributed equally: Yingguang Frank Chan, Simon Myers. ✉e-mail: robert.davies@stats.ox.ac.uk

NATURE GENETICS TECHNICAL REPORT

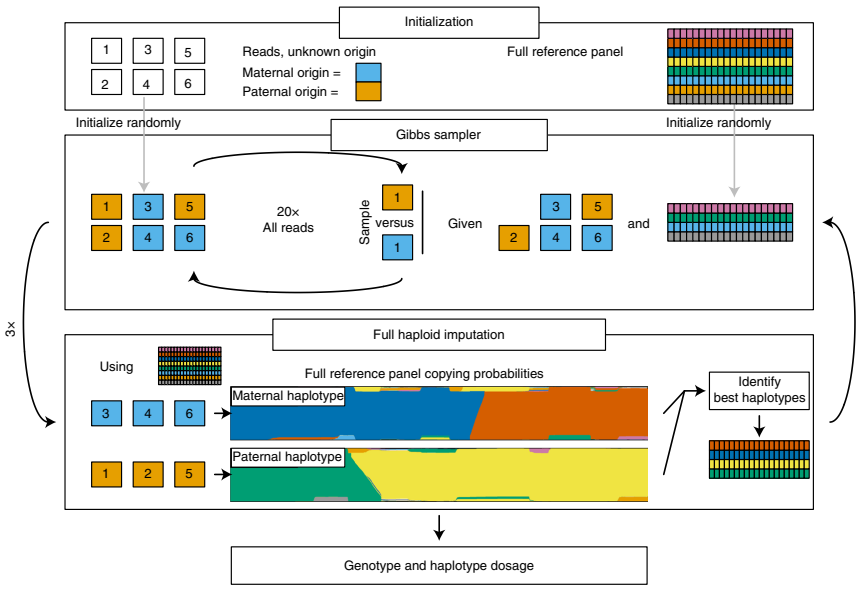



**Fig. 1 | Schematic of the QUILT model.** The model for one Gibbs sampling is shown. The model is initialized for a vector of read labels and a subset of reference haplotypes. The QUILT model then iteratively proceeds between Gibbs sampling, to obtain new read labels given the current subset of reference haplotypes, and full haploid imputation, to obtain new reference haplotype subsets using the current read labels. QUILT completes after a prespecified number of iterations. Genotype dosage is taken as an average across Gibbs samplings, while phase is taken from an additional Gibbs sampling using read labels taken as an average across previous samplings.

short-read (linked-read) sequencing, featuring partial sequencing of long (>100 kilobases (kb)) haplotype segments. We evaluate QUILT's ability to accurately partition reads and impute genotypes across these different data types using one individual sequenced on all three platforms (NA12878) and three larger additional datasets. We further demonstrate QUILT's effectiveness in imputing across diverse populations from the 1000 Genomes Project and QUILT's ability to accurately impute human leukocyte antigen (HLA) class I and II alleles. Finally, we investigate what these results mean for the relative merits of designing studies using lcWGS versus genotyping microarray technologies.

## Results

**Model overview.** Our method, QUILT, models each sample haplotype to be imputed as a mosaic of a panel of reference haplotypes using the Li and Stephens model[6]. Let $K$ be the number of reference haplotypes. Because samples are diploid, a naïve implementation of this model has computational complexity that is proportional to $K^2$, which is computationally prohibitive for large $K$. However, fundamentally, each sequencing read comes from either the maternal or paternal chromosome. Conceptually, if we could split them into two such sets, the imputation becomes computationally linear in $K$. Previously, in the STITCH model, we introduced a 'pseudo-haploid' approach that had linear computational complexity[17] by assigning approximating probabilities for each read to come from each background. In this study, we improved this approach by implementing a Gibbs sampler through efficient reuse of stored forward–backward probabilities, allowing for resampling of all read labels into the two parental chromosomes using a single forward–backward pass of the hidden Markov model (HMM). Since Gibbs sampling is slow when using the full reference panel, we iterated between using the Gibbs sampler on a subset of the reference panel to update the read labels, and performing haploid imputation using the read labels and the full reference panel to update the subset of the reference panel, and in the terminal iteration, perform the imputation (Fig. 1). By using a fraction of the panel, Gibbs sampling is fast; it is also accurate since it uses the best matching reference haplotypes. The final imputation results are taken as an average across multiple Gibbs processes and the entire algorithm is computationally linear in $K$. Further details, including the phasing, computational efficiencies and speedups used in the model, are described in the Methods and Supplementary Note.

**Imputation performance for NA12878.** We obtained a high-coverage sequence for NA12878 using standard Illumina short-read (Illumina)[24], haplotagged Illumina short-read (HT) and long-read (ONT) sequences[25]. We inferred accurate genotypes and haplotypes using high-coverage trio data and phasing both using trio information and a haplotype reference panel (Methods); we used these as approximate truth haplotypes, from which we further generated probabilities that each sequencing read ultimately came from either the maternal or paternal haplotype (Methods). In what follows, we generally imputed 200 mega bp (Mbp) of the genome for comparisons; for a reference panel, we used a subset of the HRC dataset with NA12878 removed (therefore using 54,328 haplotypes)[7] (Methods). In later sections, we use different subsets of the HRC after removing exact matches to test sequences. We chose parameter settings for QUILT that gave a reasonable trade-off between accuracy and speed across a wide range of coverages (Supplementary Fig. 1).

We assessed phasing accuracy both visually by comparing the continuity of inferred to truth read labels and quantitatively using the phase switch error rate. QUILT can achieve increasingly

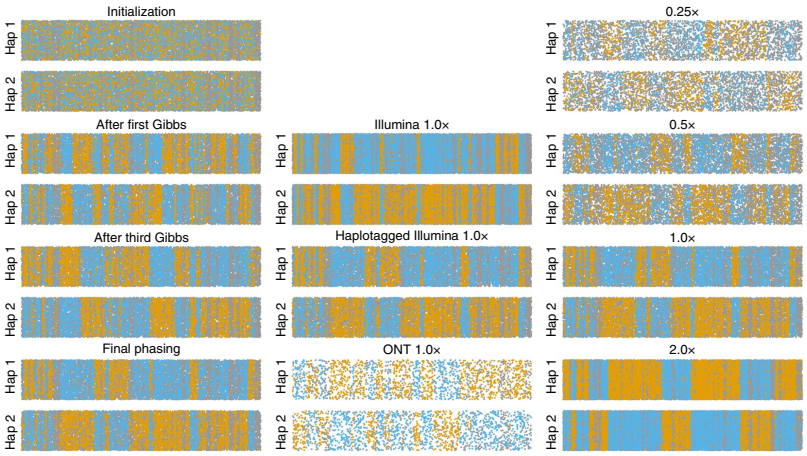



**Fig. 2 | Assessment of read label partitioning.** Per analysis, reads were grouped based on assignment to Hap1 or Hap2, with the remaining *y* axis variation being jitter. The *x* axis gives the central location of the read along 20 Mbp of chromosome 20. Reads are colored blue and orange to reflect a high posterior probability of coming from a truth maternal or paternal chromosome, while gray indicates equally likely from either truth chromosome. Switches between runs of orange and blue denote probable switch errors. The columns denote the effect of multiple iterations (leftmost for haplotagged 1.0×), different technologies (center, for 1.0×) and coverages (rightmost for haplotagged).

accurate read label phasing as the algorithm progresses, eventually achieving a read level phasing with approximately one switch error per Mb (Fig. 2; errors are visualized as flips between runs of blue and orange, over a continuous 20-Mbp region of chromosome 20). When comparing across technology platforms, standard short-read Illumina data (unlinked) generated more switches than longer-read HT or ONT data (Fig. 2). Indeed phase switch error rates were 0.09% for ONT, 0.08% for HT and 0.13% for Illumina at 1.0× coverage (Supplementary Table 1). Phasing accuracy was robustly high across a wide spectrum of sequence coverages, increasing only to 0.11% (HT) at 0.25× coverage, although genotyping error rates increased as coverage decreased (5.26% of heterozygous sites at 0.25× versus 2.12% at 1.0× were erroneously imputed as homozygous).

We next assessed genotype imputation accuracy across data types, methods and coverage (Fig. 3 and Supplementary Table 2). To explore the contribution of phasing errors to genotyping errors, we supplied our 'truth' read label origins to the QUILT framework to generate what we call 'optimal' imputation results (these are optimal in the sense of possessing idealized phase information but still vary with read coverage). We stratified results by allele frequencies from the separate Genome Aggregation Database (gnomAD)[26]. Throughout the results, we use 'rare' to refer to SNPs with 0.1–0.2% allele frequencies and 'common' to refer to SNPs with 20–50% frequencies. Encouragingly, the imputation results for QUILT at 0.5× were quite close to the optimal results and showed a benefit for HT, particularly at rare SNPs (optimal, HT and Illumina, respectively; rare $r^2$=0.76, 0.709, 0.678; common $r^2$=0.988, 0.980, 0.975). This benefit for HT remained at lower coverages, for example, for rare SNPs at 0.1× (HT and Illumina, respectively; $r^2$=0.460, 0.416). Comparisons with ONT were complicated slightly by the higher error rate of this platform; at 0.5× ONT, it showed similar accuracy to Illumina at rare SNPs (rare $r^2$=0.669) and somewhat worse accuracy for common SNPs ($r^2$=0.937), indicating a trade-off between better phasing and lower accuracy of individual reads.

We compared the QUILT results to GLIMPSE[20] for each analysis (Fig. 3a and Supplementary Table 2). Since GLIMPSE uses genotype likelihoods from a variant call format (VCF) for input, there is no gain in using linked-read HT data under GLIMPSE versus standard Illumina sequencing. At 0.5×, the results for QUILT and GLIMPSE were similar, with QUILT HT being the most accurate (QUILT HT, QUILT Illumina and GLIMPSE Illumina, respectively; rare $r^2$=0.709, 0.678, 0.672; common $r^2$=0.980, 0.975, 0.974). For ONT, accuracy was considerably lower with GLIMPSE than QUILT (QUILT ONT, GLIMPSE ONT; rare $r^2$=0.669, 0.428; common $r^2$=0.937, 0.771). Run time comparisons between both QUILT and GLIMPSE for a variety of reference panel sizes for both low (NA12878, $n$=3) and moderate (1000 Genomes Project, $n$=93) sample sizes confirmed that QUILT has approximately linear computational complexity in reference panel size while GLIMPSE has constant computational complexity in reference panel size for moderate sample sizes (Supplementary Fig. 2). Both methods had low and approximately linear relationships between memory usage and reference panel size. Both QUILT and GLIMPSE include parameter settings that trade off accuracy versus run time. Therefore we reran both methods across a range of parameter values varying from their defaults for NA12878 (Supplementary Fig. 3; $n$=3 with three data types) and Illumina for 1000 Genomes Project data (Supplementary Fig. 4; $n$=93, a more realistic scenario involving more samples). The results indicate (when considering both speed and accuracy) that for lower-coverage data (0.1×), QUILT is favored over GLIMPSE for all three data types but especially HT and ONT data. For moderate coverages (0.5×), the methods are similar except for ONT; for high-coverage data (1.0×, 2.0×), GLIMPSE is favored over QUILT for the Illumina and HT data types, while QUILT is clearly favored for ONT data.

We next compared the QUILT results to those obtainable from genotyping microarrays. We approximated microarray input using high-coverage WGS restricted to the relevant array SNP site list. We examined performance versus the commonly used Affymetrix UK Biobank (UKBB) and Illumina Global Screening Array (GSA) arrays, imputing using Beagle v.5.1 (ref. [8]) (Fig. 3b and Supplementary Table 3). Both arrays yield similar accuracies at both rare (UKBB and GSA, respectively; $r^2$=0.694, 0.683) and common SNPs ($r^2$=0.989, 0.984). For rare SNPs, all three sequencing

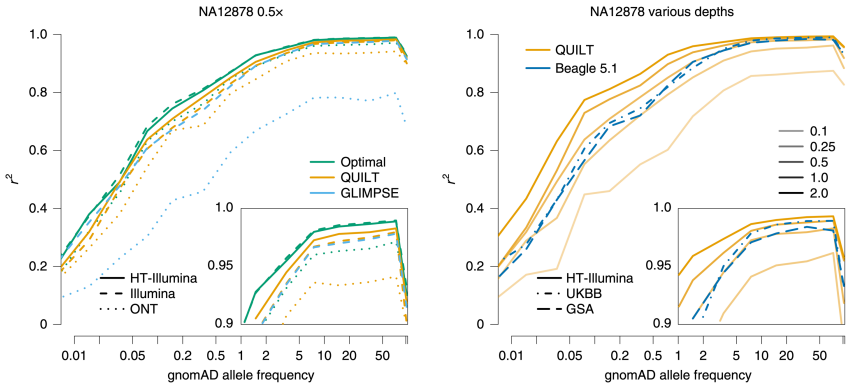



**Fig. 3 | Imputation accuracy of the NA12878 sample.** The $r^2$ per bin was aggregated over SNPs with a given gnomAD allele frequency for a given technology, coverage and method.

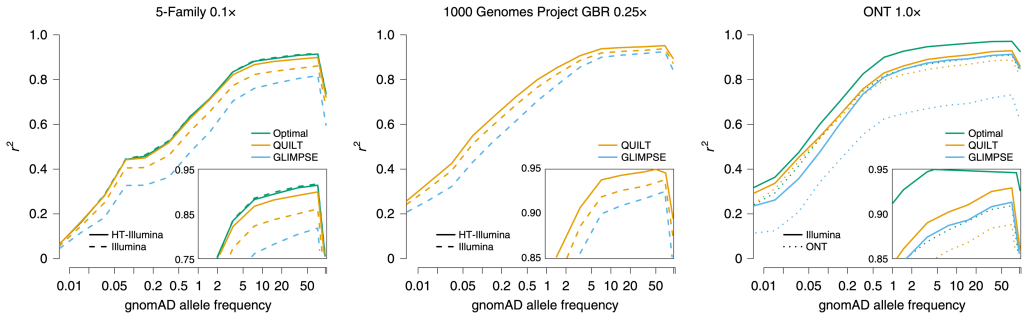



**Fig. 4 | Imputation accuracy of 5-Family, GBR and ONT samples.** The $r^2$ per bin was aggregated over all SNPs in that gnomAD allele frequency bin across all samples, for a given technology, coverage and method.

platforms already exceeded this accuracy at 1.0× coverage (HT, Illumina and ONT, respectively; $r^2$=0.776, 0.754, 0.741). For the best-performing platform (HT), even at 0.5× coverage imputation was at least as accurate as arrays ($r^2$=0.709 for rare SNPs, $r^2$=0.988 for common SNPs), demonstrating the utility of QUILT and lcWGS.

**Imputation performance across diverse sequenced samples.** In three additional datasets, we confirmed the overall patterns we saw using NA12878 between data type and imputation performance. First, we further compared HT and Illumina using 7 offspring from 5 different families of North American ($n$=3) and Ashkenazi Jewish ($n$=2) background (Fig. 4; see Supplementary Table 4 for all coverages). At 0.1×, we saw the same ordering as with NA12878 for optimal, QUILT HT, QUILT Illumina and GLIMPSE for both rare ($r^2$=0.463, 0.450, 0.407, 0.328) and common ($r^2$=0.911, 0.891, 0.850, 0.804) SNPs, although as coverage increased, the approaches became more similar. Second, we again compared HT and Illumina data using 59 samples of British background from the 1000 Genomes Project (GBR, using the 1000 Genomes Project[24] notation) (Fig. 4; see Supplementary Table 5 for all coverages). At 0.25×, we saw the same ordering for QUILT HT, QUILT Illumina and GLIMPSE for both rare ($r^2$=0.629, 0.593, 0.519, respectively) and common ($r^2$=0.947, 0.931, 0.917, respectively) SNPs. Finally, we further compared Illumina and ONT using seven samples of diverse genetic backgrounds from Shafin et al.[27] (Fig. 4; see Supplementary Table 6 for all coverages). At 1.0×, as for the NA12878 imputation, we saw that imputation from Illumina was more accurate than ONT, that QUILT was more accurate than GLIMPSE on ONT data; similarly, within data types, we saw the same relationship between optimal, QUILT and GLIMPSE both for Illumina (rare $r^2$=0.7, 0.64, 0.598; common $r^2$=0.97, 0.925, 0.908, respectively) as well as for ONT (rare $r^2$=0.638, 0.629, 0.439; common $r^2$=0.905, 0.884, 0.718, respectively).

**Imputation performance relative to genotyping arrays.** We next evaluated imputation performance versus genotyping microarrays (UKBB and GSA) across diverse samples using QUILT 1000 Genomes Project samples for five groups (ASW, CEU, CHB, PJL and PUR) from distinct continental populations (Methods)[24]. As expected, imputation accuracy was highest for CEU (Northern and Western European ancestry) samples (Fig. 5 and Supplementary Table 7), who are most similar to the bulk of the HRC reference panel, where for rare SNPs, arrays were comparable to 0.25× sequencing data imputed using QUILT ($r^2$=0.63–0.66). Interestingly for other groups, although absolute imputation quality declined, the relative performance of lcWGS and QUILT increased compared to

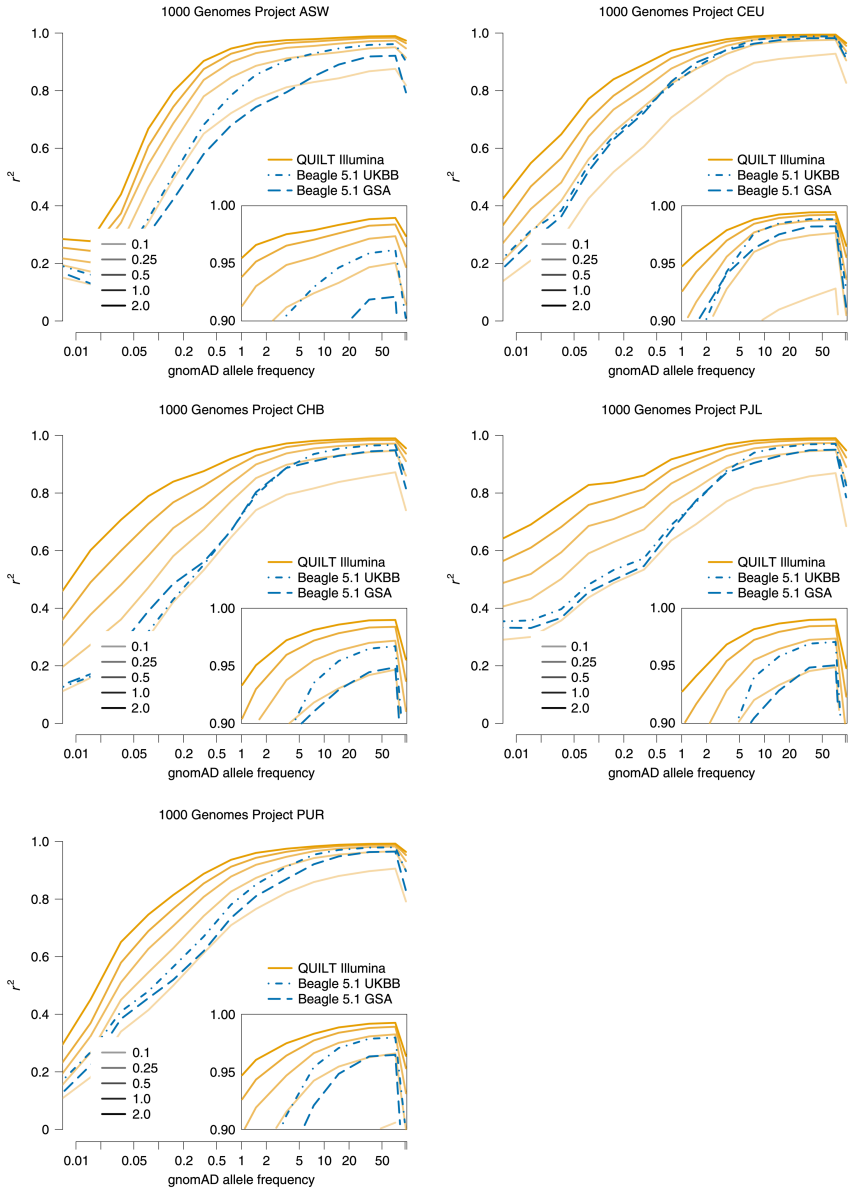



**Fig. 5 | Imputation accuracy of the 1000 Genomes Project samples.** The $r^2$ per bin was aggregated over all SNPs in that gnomAD allele frequency bin across all samples, for a given technology, coverage and method.

genotyping arrays. For example, for CHB (Han Chinese in Beijing) samples, lcWGS and QUILT outperformed the arrays for rare SNPs (QUILT, GSA and UKBB, respectively; $r^2 = 0.581$, 0.485, 0.43). For common SNPs, imputation accuracy was generally high across platforms. Thus, the benefits of sequencing versus genotyping appear considerably greater for populations less similar to a given reference dataset. Higher coverage 1.0× and 2.0× lcWGS data outperformed arrays at all frequencies, especially lower-frequency variants, with QUILT nearly doubling the above array-based values (for example, for rare SNPs, CHB, QUILT 1.0× $r^2 = 0.768$, 2.0× $r^2 = 0.840$).

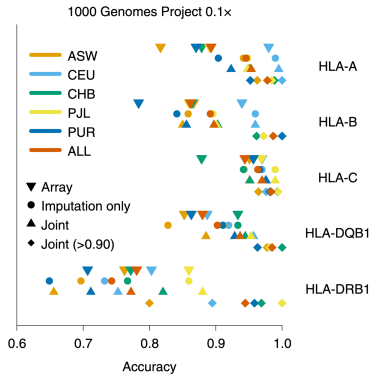



**Fig. 6 | Imputation accuracy of HLA loci.** Accuracy is the percentage of correct unphased HLA alleles versus computationally inferred truth. Results are shown both per population and in aggregate (ALL). Results are given both using only imputation as well as imputation plus direct read mapping (joint, the default QUILT output). Results are further given at the subset of individuals with confidently inferred alleles (joint (>0.90)). As reported elsewhere[29], HLA class I loci (HLA-A, HLA-B and HLA-C) are less diverse than class II loci (HLA-DRB1 and HLA-DQB1) and thus yield more accurate imputation results.

**HLA imputation performance.** Methods for accurate HLA imputation using genotypes[28,29] or high-coverage sequence data[30] have previously been developed. As part of QUILT, we developed a new HLA imputation algorithm for lcWGS that uses reads inside an HLA locus for direct read mapping and uses the remaining reads for imputation using a labeled reference panel (Methods). We used reference HLA data from both the HLA database IPD-IMGT/HLA[31] and 1000 Genomes Project and used withheld 1000 Genomes Project data for testing. We imputed the HLA types for five populations (ASW, CEU, CHB, PJL and PUR) at each of five classical HLA loci (HLA-A, HLA-B, HLA-C, HLA-DQB1, HLA-DRB1). In addition, we assessed the accuracy of array-based HLA imputation using the same reference data and a method derived from the approach of SNP2HLA (Methods)[28,32].

The results demonstrate that accurate HLA imputation from low-coverage sequence data is achievable across both class I (HLA-A, HLA-B and HLA-C) and II (HLA-DQB1 and HLA-DRB1) loci, with the former showing generally higher accuracies (Fig. 6 and Supplementary Table 8). Using HLA-A as an example, the array-based approach had an accuracy of 0.893 across all populations, while at 0.1× QUILT achieved 0.953 (0.995 CEU, 0.950 ASW (Americans of African Ancestry in SW USA)). Accuracy rose to 0.978 when only considering confidently called alleles. Accuracy decreased to 0.946 when the direct read information was not used, indicating that even at 0.1× direct read mapping can be useful. Results were consistent across coverages (Supplementary Fig. 5), although direct read mapping provided a relatively larger boost in accuracy at higher coverage (9.3% at 2.0×) versus lower coverage (1.9% at 0.1×). Results for the other class I loci of HLA-B and HLA-C, as well as HLA-DQB1, were similar, although HLA-B was less accurate across all samples, again mirroring known results[29]. The accuracies for HLA-DRB1 were lower and this was the only locus where array-based accuracy (0.781) slightly exceeded QUILT for 0.1× data (0.772). Nonetheless, accuracy at confidently called alleles remained high for QUILT at 0.1×, being 0.985 for HLA-DQB1 and at least 0.944 for every other locus. As expected, across all analyses, imputation accuracy was generally higher for more common HLA alleles than rare alleles (Supplementary Table 8). Finally, we examined the results for seven specific HLA alleles chosen because they are either disease-associated or associated with adverse drug reactions, similarly to previous work[29]. Except for some DRB alleles, accuracy at these medically important alleles was generally very high (Supplementary Table 9).

**Relative effective sample size and power.** Finally, we performed a cost–benefit analysis of lcWGS and QUILT-based imputation versus widely used genotyping microarrays. We assessed the benefit in both a GWAS and burden test setting, where for the latter the focus is on testing for an excess of rare coding variants in cases at a specific locus (Methods). We tested a scenario of using samples drawn from the CHB population, using estimates of imputation accuracy from the 1000 Genomes Project CHB analysis, and assumed a fixed cost of £30 per array and library costs from Meier et al.[16]. We varied both the cost of phenotyping a sample (that is, non-genotyping costs) and the per-× sequencing costs from approximate costs available today of US$1,000/30× genome, to potential lower future costs of US$250/30× genome.

Results showed that for both settings, sequencing yields nearly uniformly larger effective sample sizes or greater power than genotyping arrays (Fig. 7, Supplementary Table 10 and Supplementary Fig. 6). For GWAS testing, lcWGS and QUILT yield effective sample sizes 3.5 and 1.4 times larger than using genotyping microarrays for inexpensive and expensive phenotyping costs, respectively, to identify associations with rare SNPs of 0.1–0.2% frequency. For the burden test setting, power differences were even more pronounced. The relative gain in effective sample size or power increases as sequencing and phenotyping costs decrease. Results for fixed phenotyping and sequencing costs show that the improvement of using lcWGS and QUILT is especially pronounced at the lowest coverages. At current sequencing costs, the most cost-effective sequencing coverage is low, at less than 1× coverage, although imputation using higher coverages will be more useful in the future as sequencing costs decrease.

## Discussion

Inexpensive and accurate genotyping solutions are needed to perform the next generation of GWAS. In this work, we showed that QUILT can unlock the power of lcWGS for this task, simultaneously improving accuracy and reducing costs for diverse data types and individuals from varied populations, compared to traditional genotyping arrays. In this section, we consider why we observed these results, the broad implications of the results observed for GWAS and other studies and finally the natural extensions of this work.

To begin, it is helpful to consider imputation from lcWGS, supposing we knew the true read label assignment, that is, whether each sequencing read came from the maternal or paternal chromosome. Neglecting the impact of errors at rates <1% for either kind of data, standard arrays type on the order of a million SNPs across the genome, typically favoring common variants and/or higher information content. For comparison, lcWGS sequencing at 2.0× coverage—if we know the true read label assignment—yields about 1.0× coverage for each haplotype or about 63% of the entire genome for each haplotype under a simple Poisson model of sequencing coverage. While information from some array SNPs is lost, information from many more non-array SNPs is gained, explaining why QUILT can be more accurate than arrays at higher coverage, particularly for populations other than those for which genotyping microarrays are principally designed for. The above relies on accurate read label assignments; in practice, while QUILT makes some errors, assignment is highly accurate, which explains why QUILT results are often near optimal values, where direct phase information is provided. This explains why short-read (Illumina-based) sequencing outperforms long-read data (exemplified by ONT) because the

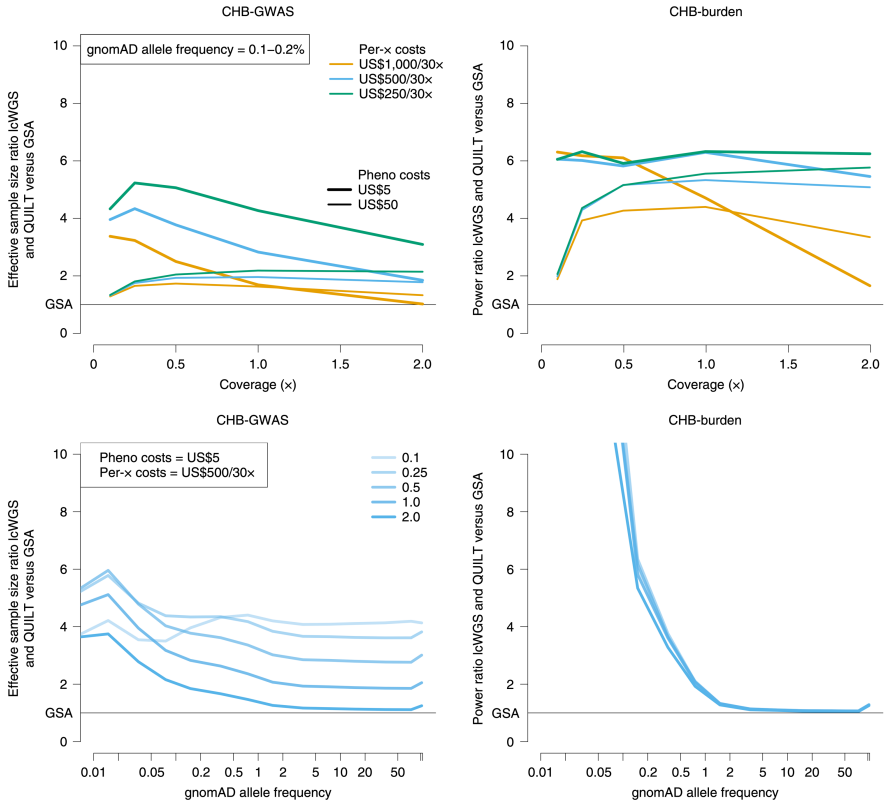



**Fig. 7 | Relative increase in effective sample size and power using lcWGS and QUILT.** Results are shown as a ratio of effective sample size for the GWAS setting and a ratio of power for the burden test setting. Results use the 1000 Genomes Project CHB imputation accuracy. Top, Results are given as a function of coverage, with variable phenotyping and per-x sequencing costs, for a fixed allele frequency (0.1–0.2%). Bottom, Results are given as a function of allele frequency, with varying coverage, assuming fixed phenotyping (US$5 per sample) and per-x sequencing costs (US$500 per 30x). All results assume a library preparation cost of £1.36 per sample and an array cost of £30 per sample.

slightly improved read label assignment possible from long reads is outweighed by the much lower per-base error rate of the short-read data. The haplotagging approach, meanwhile, improves accuracy versus unlinked Illumina reads since it retains the same per-base error rate but decreases phasing errors, particularly at low coverages, where more phasing uncertainty remains.

What do these results mean for GWAS today and in the future? lcWGS already provides similar accuracy to genotyping microarrays at only 0.1–0.25× coverage, and much higher accuracy at 1.0–2.0×. Combined with estimates of library and sequencing costs, in terms of power for a given expenditure, it is already preferable to use lcWGS and QUILT rather than genotyping microarrays across a diverse range of scenarios. This is particularly true for discovering rare disease-associated variants, which recent work suggested harbor a disproportionate amount of heritability for human phenotypes[33,34]. Because QUILT, in contrast to other methods for lcWGS, models reads directly and can work with high per-base error rates, it is robust across sequencing technologies, thereby ensuring that accuracy remains high regardless of future technological advances. Furthermore, by modeling reads directly, QUILT should work well regardless of the diversity of the region of the genome, or in nonhuman species, in which heterozygosity levels are often higher and can be an order of magnitude higher than that in most human populations[16]. The strong performance of QUILT for HLA imputation, across diverse sample backgrounds, offers a concrete confirmation of this prediction. QUILT has linear computational complexity in sample size and haplotype reference panel size, ensuring reasonable run times across a range of study sizes.

In terms of future development of the model, as datasets with millions of lcWGS become available, they will provide extensive information to improve imputation; we intend to address this in future work. For instance, we might iterate imputation, followed by incorporation of imputed sequenced samples into the reference panel. This might offer particular improvements, for example, for newly discovered rare variants not represented in the reference panel. The QUILT model could also be adapted to new settings. In one example, noninvasive prenatal testing using low-coverage sequencing is rapidly becoming the clinical standard of care[35] and can generate the genotypes, both fetal and maternal, needed for GWAS[36]. The model presented in this study could readily be adapted to this data type to

allow separate imputation of the three haplotypes present in noninvasive prenatal testing: the maternal transmitted and untransmitted and the paternal transmitted sequences. This would enable recovery of the maternal, fetal and partial paternal genomes, allowing the determination of, for example, carrier status for genetic risk variants and polygenic risk scores.

In conclusion, lcWGS imputation using QUILT is flexible and accurate across a range of coverages and data types. It is particularly beneficial for imputing rare SNPs and for imputing genotypes for human populations poorly represented by existing large haplotype reference panels; it should help empower a new wave of large, ethnically diverse GWAS.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00877-0.

Received: 14 July 2020; Accepted: 23 April 2021;
Published online: 3 June 2021

## References

1. Pasaniuc, B. & Price, A. L. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.* **18**, 117–127 (2017).
2. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
3. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590 (2018).
4. Burton, P. R. et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
5. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
6. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
7. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
8. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
9. O'Connell, J. et al. Haplotype estimation for biobank-scale data sets. *Nat. Genet.* **48**, 817–820 (2016).
10. Loh, P.-R., Palamara, P. F. & Price, A. L. Fast and accurate long-range phasing in a UK Biobank cohort. *Nat. Genet.* **48**, 811–816 (2016).
11. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
12. Pasaniuc, B. et al. Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).
13. Cai, N. et al. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* **523**, 588–591 (2015).
14. Nicod, J. et al. Genome-wide association of multiple complex traits in outbred mice by ultra-low-coverage sequencing. *Nat. Genet.* **48**, 912–918 (2016).
15. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
16. Meier, J. I. et al. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc. Natl. Acad. Sci. USA* https://doi.org/10.1073/pnas.2015005118 (2021).
17. Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969 (2016).
18. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
19. Spiliopoulou, A., Colombo, M., Orchard, P., Agakov, F. & McKeigue, P. GeneImp: fast imputation to large reference panels using genotype likelihoods from ultralow coverage sequencing. *Genetics* **206**, 91–104 (2017).
20. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
21. VanRaden, P. M., Sun, C. & O'Connell, J. R. Fast imputation using medium or low-coverage sequence data. *BMC Genet.* **16**, 82 (2015).
22. Ros-Freixedes, R. et al. Accuracy of whole-genome sequence imputation using hybrid peeling in large pedigreed livestock populations. *Genet. Sel. Evol.* **52**, 17 (2020).
23. Zheng, C., Boer, M. P. & van Eeuwijk, F. A. Accurate genotype imputation in multiparental populations from low-coverage sequence. *Genetics* **210**, 71–82 (2018).
24. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
25. Bowden, R. et al. Sequencing of human genomes with nanopore technology. *Nat. Commun.* **10**, 1869 (2019).
26. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
27. Shafin, K. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat. Biotechnol.* **38**, 1044–1053 (2020).
28. Jia, X. et al. Imputing amino acid polymorphisms in human leukocyte antigens. *PLoS ONE* **8**, e64683 (2013).
29. Karnes, J. H. et al. Comparison of HLA allelic imputation programs. *PLoS ONE* **12**, e0172444 (2017).
30. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
31. Robinson, J. et al. IPD-IMGT/HLA Database. *Nucleic Acids Res.* **48**, D948–D955 (2020).
32. Luo, Y. et al. A high-resolution HLA reference panel capturing global population diversity enables multi-ethnic fine-mapping in HIV host response. Preprint at *medRxiv* https://doi.org/10.1101/2020.07.16.20155606 (2020).
33. Durvasula, A. & Lohmueller, K. E. Negative selection on complex traits limits phenotype prediction accuracy between populations. *Am. J. Hum. Genet.* **108**, 620–631 (2021).
34. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at *bioRxiv* https://doi.org/10.1101/588020 (2019).
35. Snyder, M. W. et al. Copy-number variation and false positive prenatal aneuploidy screening results. *N. Engl. J. Med.* **372**, 1639–1645 (2015).
36. Liu, S. et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**, 347–359.e14 (2018).

## Methods

**QUILT.** In the QUILT model, imputation is performed through an iterative two-step process. First, sequencing read labels, reflecting maternal or paternal origin, are updated using Gibbs sampling based on a subset of the full reference panel. Second, based on the read labels, sequencing reads are split into two sets and separately imputed using the full reference panel; this is used both to determine the best matching haplotypes for the next round of the Gibbs sampler and, in the terminal iteration, to provide the imputed genotypes to be output. In this article, we provide higher-level detail on how this works; full details are shown in the Supplementary Note. The full details include an explanation of a generative model under which reads would be simulated, detailed mathematics for both the Gibbs sampler and full haploid imputation, a description of the phasing procedure and an explanation of other parameter choices. Note that throughout the description of the model, the term 'read' refers to potentially discontinuous sequencing information from a single DNA fragment, either a random or observed variable, but known to come from the same underlying molecule: that is, two reads from a read pair will be called a read since they are observations from the same molecule.

Let Gen be the genotype for some arbitrary SNP, and $E[\text{Gen}|O]$ be the expected genotype given the observed sequencing data. Let $\lambda$ be the parameters of the model, for example, the recombination rate. We can use random sampling to estimate the diploid genotype dosage as:

$$E\left[\text{Gen}|O,\ \lambda\right] = \sum_{g=0}^{2} \sum_{H\sim P(H|O,\lambda)} g \times P(\text{Gen} = g|H, O, \lambda)$$

We used Gibbs samplings to generate draws from $H \sim P(H \mid O, \lambda)$. Let the vector $h$ be our realized Gibbs sampled value at any point during the sampler. We begin by initializing $h$ at random, that is, for the read indexed by $v$, $h_v$ is drawn equally from {1,2} for all $v$ (where arbitrarily 1 = maternal, 2 = paternal). Let $o$ be the realized observations for the random variable $O$ (sequencing read bases and base qualities). We assume that the bases of a sequence read reflect the haplotype being copied from the haplotype reference panel—specifically, who is being copied from at the central point of the read. Suppose we consider a Li and Stephens[6] model and an HMM, where transition probabilities depend on the recombination rate in the usual way and emission probabilities for a state (copied reference haplotype) at a site depend on the reference haplotype sequence and observed sequencing reads whose central location is that site. (By the assumption above, the observations (reads) are independent between sites, ensuring this is a valid HMM.) From this, we can calculate $P(O=o \mid H=h, \lambda)$ using the forward–backward algorithm. Now, suppose we want to sample a new value at read label $h_v$ for read $v$ conditional on all other read labels. Let $H_v$ be this random variable at this point in the Gibbs sampler and let $H_{-v}$ be a random variable representing the remaining read labels. We therefore need to calculate:

$$P\left(H_v = l|H_{-v} = h_{-v},\ O = o, \lambda\right) = \frac{P(O = o, H_v = l,\ H_{-v} = h_{-l}|\lambda)}{\sum_{b=1}^{2} P(O = o, H_v = b,\ H_{-v} = h_{-v}|\lambda)}$$

for l in {1,2} and sample $h_v$ using this probability (where we note $P(O=o, H=h \mid \lambda) = P(O=o \mid H=h, \lambda)\ P(H=h \mid \lambda)$ is easy to switch between as $P(H=h \mid \lambda)$, the probability of the read labels, is 1/2 to the power of the number of reads). Naively calculating the above requires a new forward–backward pass of the HMM. We can avoid a new forward–backward pass through efficient reuse of the forward and backward variables and in fact resample all read labels under the Gibbs sampler using a single forward–backward pass. Details of this, and further details including a block Gibbs sampler, are given in the Supplementary Note.

Now, while the above is computationally linear in $K$, the reference panel size, it can be slow for large $K$. Therefore, we ran the above Gibbs sampler using a reduced set of haplotypes (default 400). Furthermore, the above HMM is based on the assumption that each read has a central location and genotypes in sequencing reads are based on the reference haplotype copied at that central location. This assumption is less accurate for long-read sequencing; moreover, it is irrelevant once read labels are known since once reads are known to come from the same haplotype, membership of a sequenced base in a read no longer matters. This in effect changes the nature of the observations at a given site from sequencing reads to sequenced bases and their base qualities and allows us to work with haplotype genotype likelihoods. Therefore, given read labels, we split the reads into sets reflecting maternal and paternal origin, generated haplotype genotype likelihoods and performed full haploid imputation using the entire reference panel. We then used posterior state probabilities to update the subset of the reference panel used by the Gibbs sampler, as well as optionally outputting the genotype dosages. Full details of this are given in the Supplementary Note.

**Datasets.** *NA12878.* Both the NA12878 sample and the further samples from the 1000 Genomes Project center came from the New York Genome Center (NYGC) resequencing effort. We generated haplotagged NA12878 Illumina short-read data; by either considering or ignoring the BX tags, we used these data both for the analyses involving haplotagging and those without. We used approximately 80× ONT data from Bowden et al.[25] and converted it to GRCh38 by first downsampling it using SAMtools v.1.0 (ref. [37]) to approximately 8×, converting to the FASTQ file format using SAMtools, mapped with Minimap2 v.2.1 (ref. [38]) using --alt-drop 1.0 and using --alt with a list containing all nonautosomal, chromosome X, chromosome Y or chromosome MT contigs and sorted and indexed them using SAMtools v.1.10. Parental genomes were obtained from the Illumina Platinum Genomes project under the European Nucleotide Archive (ENA) accession no. PRJEB3381 and converted to GRCh38 per chromosome for chromosomes 6, 20 and 21 by first sorting reads by read name using SAMtools, then converting to FASTQ format using SAMtools, then read mapping using the Burrows–Wheeler Aligner (BWA) MEM v.0.7.15 and then sorting; then, results were combined using SAMtools cat and finally resorted and indexed.

*5-Family.* We generated 9.4–16.4× coverage (mean 12.2×) haplotagged data for imputation and truth data for 7 offspring and 10 parents in 5 families from the Bloom Syndrome Repository (4 trios and one 2-parent plus 3-offspring family). Parental samples were similarly haplotagged (2.8–9.8×; mean 6.5×), although this information was not used for the inference of 'true' genotype or phasing. Samples were North American and also included individuals of Ashkenazi Jewish origin. Written informed consent from all 5-Family individuals was obtained by the institutional review board of Weill Cornell Medicine.

*GBR.* We generated low-coverage (range 0.13–0.47×; mean 0.30×) haplotagged data for the 91 GBR 1000 Genomes Project samples as described in the haplotagging section of the Supplementary Note. We analyzed the 59 samples with assessed depth ≥0.25×. We used 'truth' data for these samples from the 1000 Genomes Project resequencing data from the NYGC[24].

*ONT.* We downloaded high-coverage ONT and 10× Illumina data from Shafin et al.[27] for seven samples[27] (HG01243, HG02080, HG03098, HG01109, HG02055, HG02723, HG03492) from Amazon's pangenomics resource (https://s3-us-west-2.amazonaws.com/human-pangenomics/). We downloaded high-coverage Illumina 1000 Genomes data for their parents from the NYGC. For the ONT data, we processed the first 20 million reads from the FASTQ file using Minimap2 v.2.1 in the same way as the NA12878 sample. For the 10× Illumina data, we used the proc10xG toolkit (https://github.com/ucdavis-bioinformatics/proc10xG; commit 7afbfcf), which successively runs process_10xReads.py, BWA MEM, samConcat2Tag.py and SAMtools view -b to generate a mapped BAM file.

*1000 Genomes Project data.* We used high-coverage 1000 Genomes Project resequencing data from the NYGC[24]. We chose one population (ASW, CEU, CHB, PJL, PUR) from each of the 5 continental superpopulations within 1,000 G (AFR, EUR, EAS, SAS, AMR). To minimize computation time for full imputation, we chose every 5th member of this subset to impute, resulting in $n = 11$ ASW, $n = 23$ CEU, $n = 19$ CHB, $n = 20$ PJL, $n = 20$ PUR samples. Due to a lack of consistent parent or offspring high-coverage genome availability, we did not phase the data. For the imputation of HLA loci, we used all members of each population and we generated and applied a modified reference panel, as described in the HLA imputation section of the Methods.

**Mapping.** All mapping was done against the human reference genome GRCh38; when performed with the BWA MEM, it used (at least) options -Y -K 100,000,000. Average depth was calculated using SAMtools depth v.1.10 -a -q 10 -Q 10 on chromosome 20 from 1 to 10 Mbp for each BAM. Reads were grouped into molecules, either using paired-read information or using the BX tag from haplotagging. Downsampling was performed per molecule (that is, all reads from both read pairs and given the linkage information from the BX tag were taken as being from the same molecule) using a Bernoulli probability determined by the desired coverage. For the 5-Family samples, we used results from an earlier version of the chemistry for haplotagging since a later rerun achieved <2× coverage for each sample. We then considered the proportions of molecules that had no BX tag, had two or fewer reads in a BX tag or had three of more reads in a BX tag. When downsampling using data from the old chemistry, we also downsampled to ensure these proportions matched the results from the new chemistry. Subsampled BAMs were written either using the original read names (QNAME field of the BAM) or using new read names that incorporated the BX molecule information.

**Reference panel.** We used a controlled access version of the HRC[7] with GenBank accession no. EGAD00001002729. We used LiftOver v.2.22.2 to convert the reference panel from the GRCh37 to the GRCh38 reference genome using Genome Analysis Toolkit (GATK) Picard LiftoverVCF v.2.22.2 (ref. [39]). From the original GRCh37 HRC reference panel (39,131,578 autosomal variants), we removed 6,803 variants due to failure to map between the GRCh37 and GRCh38 reference genomes and a further 9,010 variants due to mismatching chromosomes between the two reference genomes. The resulting autosomal GRCh38 HRC reference panel contained 39,115,765 variants and 27,165 samples with 54,330 haplotypes.

For imputation, we created three versions of the reference panel based on three different subsets of the full panel. We first made a subset where we removed only NA12878; this was used for the analyses involving NA12878 and the 5-Family individuals. Second, we made a version where we removed NA12878, the parents from the ONT samples from Shafin et al.[27] used in this study and removed those

samples from the five populations of the 1000 Genomes Project we used to test imputation (ASW, CEU, CHB, PJL, PUR); we used this for the ONT and 1000 Genomes Project analyses. Third, we made a version where we removed NA12878 and the entire GBR population for the analyses involving the GBR dataset.

**Imputation.** We used QUILT v.0.1.3 (ref. [40]), GLIMPSE[20] v.1.0.0 and Beagle v.5.1 (ref. [8]) using default parameters across all runs. All imputation was done in 2-Mbp chunks with 500 kbp flanking buffers for all methods on three regions: chromosome 6:1–172,000,000 bp; chromosome 21:1–26,000,000 bp; and chromosome 21:14,000,001–46,000,000 bp. QUILT utilizes a preconstructed compressed internal data format derived from the relevant .hap and .legend file formats. For QUILT and GLIMPSE, we used a CEU-based recombination map (CEU_omni_recombination_20130507.tar) and LiftOver to generate a GRCh38 recombination rate map across all runs. For Beagle, we used a PLINK format recombination rate map available from the Beagle website (plink.<chr>.fixed.GRCh38.map). For GLIMPSE, we used GATK v.3.8-1-0-gf15c1c3ef UnifiedGenotyper–genotyping_mode GENOTYPE_GIVEN_ALLELES–output_mode EMIT_ALL_SITES to generate genotype likelihoods at sites to impute, which was done using GLIMPSE_phase. For both QUILT and GATK as the input for GLIMPSE, we used a minimum base quality and mapping quality of ten (using -minBQ in GATK). For input to the Beagle-representing arrays, we intersected the high-coverage truth genotypes called using GATK UnifiedGenotyper with sites present on the array to generate an input VCF. To determine sites, for the UKBB Affymetrix array, we used an annotation file (Axiom_tx_v1.na35.annot.csv), removing a small number of array sites that did not map to the autosomes or sex chromosomes, were not unique or were multiallelic and used LiftOver to get results in GRCh38, yielding 761,888 SNPs. For the Illumina GSA, we used strand files from https://www.well.ox.ac.uk/~wrayner/strand/, specifically GSA-24v3-0_A2-b38.strand.

**Assessing imputation and phasing accuracy.** We called 'truth' genotypes using high-coverage Illumina short-read BAMs at HRC biallelic SNPs using the GATK UnifiedGenotyper software and option '--alleles' at the biallelic SNPs, setting genotyping_mode GENOTYPE_GIVEN_ALLELES and output_mode EMIT_ALL_SITES[39]. We used truth genotypes from high-coverage sequencing only at sites where depth was at least six. Phasing of high-coverage trios was done first assuming Mendelian inheritance, excluding triple-heterozygous sites, using bespoke R v.3.62 code; then, the excluded sites were phased with this scaffold using shapeit4 (ref. [11]) v.4.0 (conda installation) using the HRC reference panel with only NA12878 removed.

To assess genotype imputation accuracy, we used imputed (test) dosages and high-coverage (truth) genotypes. We assessed results either per sample or across samples for SNPs in a given frequency range by constructing vectors of test dosages and truth genotypes and taking their squared Pearson correlation (in R using 'cor') at pairwise complete sites (that is, ignoring truth sites with depth less than six). SNP frequencies were taken from gnomAD[26] v.3.0, while for the very small number of SNPs in HRC not in gnomAD we used their HRC allele frequencies. We downloaded gnomAD data from https://gnomad.broadinstitute.org/downloads using links like the following for chromosome 20: https://storage.googleapis.com/gnomad-public/release/3.0/vcf/genomes/gnomad.genomes.r3.0.sites.chr20.vcf.bgz.

To assess phasing accuracy, we used switch error rates as follows. First, consider that we have both 'truth' integer haplotype data and test imputed haplotype dosages (that is, two real numbers within the range of 0–1) at sites where the truth genotypes are heterozygous. Next, define as discordant any test sites that are also not heterozygous (that is, the sum of haplotype dosages does not round to 1). On the remaining sites, define a phase switch error when either the truth haplotypes record a change where the haplotype carries the alternate allele between adjacent heterozygous sites when the test haplotypes do not or vice versa. We removed from consideration sites that were flipped, that is, yielding consecutive phase switch errors. The phase switch error rate is the number of phase switch errors divided by the total number of pairs of consecutive heterozygous sites examined and can be combined across discrete imputed windows.

**HLA imputation.** We used HLA reference data built as follows. Briefly, we downloaded full-length HLA alignments for annotated HLA genes and pseudogenes from the HLA database IPD-IMGT/HLA[31] v.3.39. This provides a set of (aligned) sequenced alleles for each region to which reads can be mapped. Separately, for each HLA region, we subsetted reference haplotypes from the HRC to samples from the 1000 Genomes Project[24]. We then obtained unphased HLA types for those samples (v.20181129), previously inferred using high-coverage exome sequence data and using the PolyPheMe software[41], and previously shown to have high accuracy. We phased HLA types onto haplotypes using a bespoke approach and then excluded all members of the test populations and those individuals with unphased alleles from the labeled reference panel. For full details, see the Supplementary Note.

For HLA imputation, we used QUILT-HLA v.0.1.6 for imputation from lcWGS. For array-based imputation, we were unable to run either SNP2HLA or the SNP2HLA module of HLA-TAPAS without error, so we used a custom HLA imputation approach implementing the algorithm used by SNP2HLA[28,32]. Briefly, we first generated genotypes simulating an array using high-coverage WGS data at array sites, then used the HRC to impute additional sites from those genotypes using Beagle v.4.1 (ref. [8]) and then finally used the labeled haplotype reference panel and the imputed genotypes to impute HLA types using Beagle. For full details, see the Supplementary Note.

**Cost-effectiveness.** We assessed the relative cost benefits of lcWGS and genotyping microarrays using the imputed results for rare (0.1–0.2%) and common SNPs (20–50%) for the CHB 1000 Genomes Project imputation results. We assumed a fixed array cost of £30 per array, library cost of £1.36 (ref. [16]) and per-× sequencing costs from US$1,000, 500 and 250 per sample, that is, per-× costs of 1,000/30, 500/30 and 250/30, converting into pounds sterling by using an exchange rate of 0.79375. For the GWAS-type analysis, for a fixed budget, we calculated the number of samples available for imputation as budget/(pheno_cost + library_cost + per_X_cost × coverage), while for the genotyping microarrays, the number of samples is budget/(pheno_cost + cost_array). We then took the effective sample size as the sample size multiplied by the imputation $r^2$ and took the ratio of these to be the relative increase in effective sample size from using lcWGS.

For the burden style analysis, we assumed a gene with 10 causal SNPs each possessing a frequency of 0.1% in cases and 0.01% in controls and used the same imputation $r^2$, that is, the imputation $r^2$ for SNPs at frequency of 0.01% in the population. For simplicity, we approximated this as 1 causal SNP with a frequency of $f_1 = 1\%$ in cases and $f_2 = 0.1\%$ in controls. We then supposed that error introduction is governed by some parameter $m$, such that in cases where X is the true genotype and Y is the observed genotype (with imputation error), $P(X=0, Y=0) = 1 - f_1 - m f_1$, $P(X=0, Y=1) = m f_1$, $P(X=1, Y=0) = m f_1$, $P(X=1, Y=1) = f_1 (1-m)$, and for controls, that $P(X=0, Y=0) = 1 - f_2 - m f_1$, $P(X=0, Y=1) = m f_1$, $P(X=1, Y=0) = m f_2$, $P(X=1, Y=1) = f_2 (1-m)$. We can then calculate $m$ given $r^2$. We then estimated power using 10,000 simulations with Fisher's exact test, given an alpha of 0.05/20,000 (approximately Bonferroni correcting for the number of genes in the genome) and given equal case and control numbers governed by the same $n$ as for the GWAS analysis, where the budget was 10,000 × (30 + pheno_cost), that is, the default budget was the array cost plus the phenotyping cost.

**Reporting Summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

The HRC haplotypes are available at the European Genome-phenome Archive under accession no. EGAD00001002729; they are available through the Sanger Institute under controlled access. The high-coverage, whole-genome sequence from the 1000 Genomes NYGC collection is available at https://www.internationalgenome.org/data-portal/data-collection/30x-grch38. Specifically, we used file http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/1000G_2504_high_coverage.sequence.index. High-coverage ONT data from Bowden et al.[25] are available through the ENA under accession no. PRJEB30620. High-coverage ONT and Illumina (10×) samples from Shafin et al.[27] are available through https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=. gnomAD SNP frequencies from the version 3.0 release were downloaded as detailed at https://gnomad.broadinstitute.org/downloads from URLs such as https://storage.googleapis.com/gnomad-public/release/3.0/vcf/genomes/gnomad.genomes.r3.0.sites.chr1.vcf.bgz. IPD-IMGT/HLA data were downloaded through their github database (https://github.com/ANHIG/IMGTHLA), specifically v.3.39 through https://github.com/ANHIG/IMGTHLA/blob/032815608e6312b595b4aaf9904d5b4c189dd6dc/Alignments_Rel_3390.zip?raw=true. Previously inferred HLA types for 1000 Genomes Project participants (v.20181129) were downloaded from the 1000 Genomes FTP (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HLA_types/20181129_HLA_types_full_1000_Genomes_Project_panel.txt). Recombination rates for the CEU 1000 Genomes Project samples were downloaded from ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/working/20130507_omni_recombination_rates/CEU_omni_recombination_20130507.tar. All new high- and low-coverage sequencing done for this study are available at the Sequence Read Archive under BioProject accession no. PRJNA669554.

## Code availability

QUILT is available from https://github.com/rwdavies/QUILT under a General Public License. The specific versions of QUILT used in this manuscript are available from Figshare[40].

## References

37. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
39. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).

40. Davies, R. QUILT source code from manuscript. *figshare* https://doi.org/10.6084/m9.figshare.14401904.v1 (2021).
41. Abi-Rached, L. et al. Immune diversity sheds light on missing variation in worldwide genetic diversity panels. *PLoS ONE* **13**, e0206512 (2018).

## Acknowledgements
We thank C. Lanz, R. Schwab, O. Weichenrieder and I. Bezrukov at the MPI Developmental Biology for assistance with high-throughput sequencing and associated data processing and A. Noll and the MPI Tübingen IT team for computational support. We used high-coverage resequencing of 1000 Genomes Project data performed by the NYGC. These data were generated at the NYGC with funds provided by National Human Genome Research Institute grant no. 3UM1HG008901-03S1. The research was supported by the Wellcome Trust Core Award Grant no. 203141/Z/16/Z with additional support from the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre and by Wellcome Trust grant nos. 200186/Z/15/Z and 212284/Z/18/Z (to S.M.). The views expressed are those of the author(s) and not necessarily those of the NHS, NIHR or the Department of Health. We acknowledge the contribution and support from affected persons and their families who contributed to the Bloom Syndrome Repository. We thank the New York Community Trust and Weill Cornell Medicine's Clinical and Translational Science Center for providing funding. M.K., D.S. and Y.F.C. are supported by the Max Planck Society and a European Research Council Starting Grant (no. 639096 HybridMiX).

## Author contributions
R.W.D. developed and implemented QUILT. M.K., D.S. and Y.F.C. developed haplotagging. R.W.D., M.K. and S.S. performed the analyses. M.F. and C.M.C. developed the 5-Family dataset. S.M. developed and implemented the QUILT-HLA typer. R.W.D., Y.F.C. and S.M. wrote the paper. All authors reviewed and approved the final manuscript.

## Competing interests
M.K. and Y.F.C. declare competing interests in the form of patent and employment by the Max Planck Society. The remaining authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00877-0.

**Correspondence and requests for materials** should be addressed to R.W.D.

**Peer review information** *Nature Genetics* thanks Sayantan Das and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at www.nature.com/reprints.

# List of figures