bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Chromosome breaks in breast cancers occur near herpes tumor virus sequences and explain why cancer drugs stop working

Bernard Friedenson

Dept. of Biochemistry and Molecular Genetics, College of Medicine
University of Illinois Chicago; bernief@UIC.edu

Correspondence: bernief@UIC.edu

Abstract

DNA matching the human Epstein-Barr virus (EBV or Human Herpes Virus 4) consistently surrounds breast cancer breakpoints, which are often near known EBV binding sites. Nearly 2000 breast cancers were compared to known EBV related cancers using publicly available data. Breast cancer breakpoints on all chromosomes cluster around the same positions as in nasopharyngeal cancers (NPC), epithelial cell cancers that are 100 per cent associated with EBV and eruptions of viral mutations. Breast cancer breakpoints also gather at the same differentially methylated regions as in NPC. Breast cancer further shares with other cancers an EBV methylation signature that inactivates complement. Characteristic breakpoints in Burkitt's lymphoma (another model cancer caused by EBV) are also near EBV matching sequences and clusters of breast cancer breakpoints. Evidence supports EBV guiding abnormal rejoining of chromosome fragments in breakage-fusion-bridge cycles. Cell division then causes more chromosome breaks and more driver mutations. This model explains why initially effective anti-driver mutation cancer drugs stop working.

Keywords: Breast cancer infection, breast cancer immunity, breast cancer virus, nasopharyngeal cancer, EBV cancer, hereditary breast cancer, BRCA1, BRCA2.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Introduction

By the time breast cancer occurs, its causes are hopelessly buried under multiple risk factors, thousands of mutations, and a lifetime of exposure to mutagens. One risk factor is Epstein-Barr virus (EBV/HHV4) which exists as a latent infection in almost all humans (>90%). Although most people control EBV infection, (Fina et al., 2001; Lawson and Glenn, 2021), EBV increases breast cancer risk by 4.75 to 6.29-fold according to meta-analyses of 16 or 10 studies, respectively. An analysis of 24 case-control studies showed EBV is significantly more prevalent in breast cancer tissues than in normal and benign controls (Lawson and Glenn, 2021). Infection occurs early, with 50% of children age 6-8 already seropositive, and seropositivity increases to 89% at age 18-19 (Balfour et al., 2013). Epidemiological associations between breast cancer and EBV infection occur in diverse geographical locations (Fina *et al.*, 2001; Peng et al., 2014; Sinclair et al., 2021).

In breast epithelial cell models, EBV infection facilitates malignant transformation and tumor formation (Hu et al., 2016). Breast cancer cells from biopsies of sporadic and hereditary breast cancers express latent EBV infection gene products (Ayee et al., 2020; Lawson and Glenn, 2021), even after excluding the possibility that the virus comes from lymphocytes (Lorenzetti et al., 2010). Evidence for the activated EBV lytic form in breast cancer predicts a worse outcome (Marrao et al., 2014). EBV can trans-activate endogenous retroviruses (Bruce et al., 2021), but unlike retroviruses, EBV does not have an integrase enzyme. Integration sequences are generally short (Xu et al., 2019b; Zapatka et al., 2020), with alternate explanations possible. However, HHV6A and 6B do integrate into human chromosomes by a mechanism thought to involve recombination (Peddu et al., 2019). Despite biological plausibility, studies testing the involvement of HHV6 in breast cancer have had methodologic limitations and are still inconclusive (Eliassen et al., 2018).

Chromosome breaks and rearrangement abnormalities in breast cells are typically more abundant in populations at high-risk for breast cancer. Even a single break in one cell can destabilize the entire human genome and generate many further complex rearrangements (Umbreit et al., 2020). In addition, EBV activation from its latent state causes massive changes in host chromatin methylation and structure (Kim et al., 2020; Tang et al., 2012). Aberrant methylation in breast cancer occurs at hundreds of host gene promoters and distal sequences (Batra et al., 2021).

The present study determines whether nearly universal human EBV infection is associated with breaks and other changes in breast cancer chromosomes. The work grew from the observation that both hereditary and sporadic breast cancers have the same kinds of chromosome damage seen in cancers with established EBV relationships. These model EBV-related cancers sometimes have wildly inappropriate chromosome interconnections at multiple different breakpoints. In lymphomas, EBV infection accompanies increased numbers of chromosome breaks, rearrangements, fusions, deletions, and insertions (Cuceu et al., 2018). EBV causes endemic Burkitt's lymphoma which provides a classic example of these rearrangements. About 85% of Burkitt's lymphomas translocate the IgH gene locus to a position 5' of MYC

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

coding regions. The translocation occurs early in tumor development and is a key driving event (Spehalski et al., 2012).

Nasopharyngeal cancer (NPC) serves as a model for an EBV-associated epithelial cell cancer because NPC has an explicit EBV connection (Germini et al., 2020; Hau et al., 2020; Xu et al., 2019a). 100% of malignant cells are EBV-positive, but the viral genome in the tumor has many single nucleotide polymorphisms. Nearly 8500 EBV forms are present in patients with EBV-associated cancers with over 2100 variants in a single host, each differing only slightly from a reference genome. Variants of the viral gene BALF2 are significantly associated with NPC (Xu *et al.*, 2019a). BALF2 is a single-strand DNA binding protein active during lytic phase (Tsurumi et al., 1996). NPC has mutations affecting innate immunity, such as those in TGF-BR2, TGF-B2, TLR3, and interferon-alpha and gamma receptors. Moreover, NF-KB pathways constitutively activate an inflammatory response (Bruce *et al.*, 2021). These changes release controls on EBV infection. In breast cancer, mutations or downregulation in DNA repair genes linked to *BRCA1-BRCA2* mediated repair pathways compromise immunity (Friedenson, 2013). Immune deficits in sporadic breast cancers are well-known.

Because of whole-genome sequencing of Burkitt's lymphomas and the epithelial cancer NPC (Bruce *et al.*, 2021), it is possible to compare breakpoints in cancers with proven relationships to herpes viruses to breakpoints in breast cancers.

Materials and Methods

Breast cancer genomic sequences

Characteristics of hereditary breast cancers compared to viral cancers. The selection of hereditary breast cancer genomes for this study required patient samples with a known, typed *BRCA1* or *BRCA2* gene mutation from two studies (Nik-Zainal et al., 2016; Nones et al., 2019). These hereditary cancers were mainly stage III ductal breast cancers or breast cancers having no specific type (Nones *et al.*, 2019). The COSMIC database curated from original publications (Nik-Zainal *et al.*, 2016; Nik-Zainal et al., 2019) allowed comparisons of hereditary to sporadic breast cancer breakpoints. 74 hereditary or likely hereditary breast cancers were selected as typed BRCA1 or BRCA2 mutation-associated cancers or breast cancers diagnosed before age 40. A study of familial breast cancers contributed another 65 BRCA1/BRCA2 associated breast cancers (Nones *et al.*, 2019). Results were checked against breakpoints in 101 triple-negative breast cancers from a population-based study (Staaf et al., 2019). Genome sequencing had been done before treatment began. Male breast cancers and cancers with BRCA1 or BRCA2 mutations diagnosed after age 49 were excluded since such mutations are less likely to be pathogenic. Cancers with hereditary mutations in PALB2 and p53 were also excluded. Sporadic breast cancers were taken as those diagnosed after age 70 in the absence of known inherited mutation.

Hereditary and sporadic breast cancer patient DNA sequence data. Gene breakpoints for inter-chromosomal and intra-chromosomal translocations were obtained from the COSMIC catalog of somatic mutations as curated from original publications or from original articles (Nik-Zainal *et al.*, 2016; Nones *et al.*, 2019; Staaf *et al.*, 2019). The GrCH38 human genome version was used, and chromosome coordinates were converted to GrCh38 when

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

necessary. DNA flanking sequences at breakpoints were downloaded primarily using the UCSC genome browser but did not differ from sequences obtained using the Ensembl genome browser. Positions of differentially methylated regions near breast cancer breakpoints (Tang *et al.*, 2012) were compared to breakpoint positions in a set of 70 NPCs based on data of Bruce et al. (Bruce *et al.*, 2021)

Fragile site sequence data. Positions of fragile sites were from a database (Kumar et al., 2019) and original publications (Maccaroni et al., 2020). The presence of repetitive di- and trinucleotides was used as a test for the exact positions of fragile sites. "RepeatAround" tested sequences surrounding breakpoints for 50 or fewer direct repeats, inverted repeats, mirror repeats, and complementary repeats.

Comparisons of DNA sequences. The NCBI BLAST program (MegaBLAST) and database (Altschul et al., 1990; Mount, 2007; Zhang et al., 2000) compared DNA sequences around breakpoints in *BRCA1*-and *BRCA2*- mutation-positive breast cancers to all available viral DNA sequences. E(expect) values <1e-10 were considered to represent significant homology. In many cases, expect values were 0 and always far below 1e-10. Virus DNA was from BLAST searches using "viruses (taxid:10239)" with homo sapiens and uncharacterized sample mixtures excluded. EBV DNA binding locations on human chromosomes were obtained from publications (Kim *et al.*, 2020; Lu et al., 2010; Xiao et al., 2016), from databases, by interpolating published figures, or by determining the location of genes within EBNA1 binding sites. EBNA1 binding data was based on lymphoblastoid and nasopharyngeal cancer cell lines. When necessary, genome coordinates were all converted to the GrCH38 version. Breaks in hereditary breast cancers were compared to EBV DNA binding sites, epigenetic marks on chromatin, genes, and copy number variations. The MIT Integrated Genome Viewer (IGV) with ENCODE data loaded and the UCSC genome browser provided locations of H3K9Me3 chromatin epigenetic modifications. The ENCODE website also provided positions of H3K9Me3 marks (www.ENCODEproject.org).

Homology among viruses was determined by the method of Needleman and Wunsch (Needleman, 1970).

Data analyses. Microsoft Excel, OriginPro, StatsDirect, Visual basic, and Python + Biopython (Cock et al., 2009) scripts provided data analysis. Excel worksheets were often imported into Python Jupyter notebooks for extended study. Chromosome annotation software was from the NCBI Genome Decoration page and the Ritchie lab using the standard algorithm for spacing (Wolfe et al., 2013).

Statistics. Statistical analyses used StatsDirect statistical software. A scatter plot of the data was always prepared before running any statistical analysis. The Fisher exact test compared viral similarities around breakpoints in hereditary breast cancers to similarities in breakpoints generated by random numbers. Tests for normality included kurtosis and skewness values and evaluation by methods of Shapiro-Francia and Shapiro-Wilk (Shapiro, 1965). Spearman tests compared distributions of the same numbers of chromosome locations that matched viral DNA. Because the comparisons require the same numbers of sites, comparisons truncated data down to a minimum value of maximum homology (human DNA vs. viral DNA) of at least 400. Excel

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

compared the positions of breast cancers vs. midpoints of genes containing repetitive DNA fragile sites on the same chromosome.

Genes associated with the immune response damaged in breast cancer. Breast cancer somatic mutations in genes were compared to genes in the immune metagenome (Charoentong et al., 2017; Lynn et al., 2008; Ortutay et al., 2007a; b). Sets of genes involved in immune responses also mediate other functions and represent a vast and growing dataset (geneontology.org). Genes involved in cancer control by immune surveillance and immunoediting are not well characterized. An extensive validation included both direct and indirect effects of gene mutations. In addition, the Online Mendelian Inheritance in Man database (www.OMIM.org) was routinely consulted to determine gene function with frequent further support obtained through PubMed, Google scholar, GeneCards, and UniProtKB. The "interferome" was also sometimes used [www. interferome.org].

Results

Viral homologies around breakpoints in *BRCA* - associated breast cancers cluster around breakpoints in NPC, a model cancer caused by EBV.

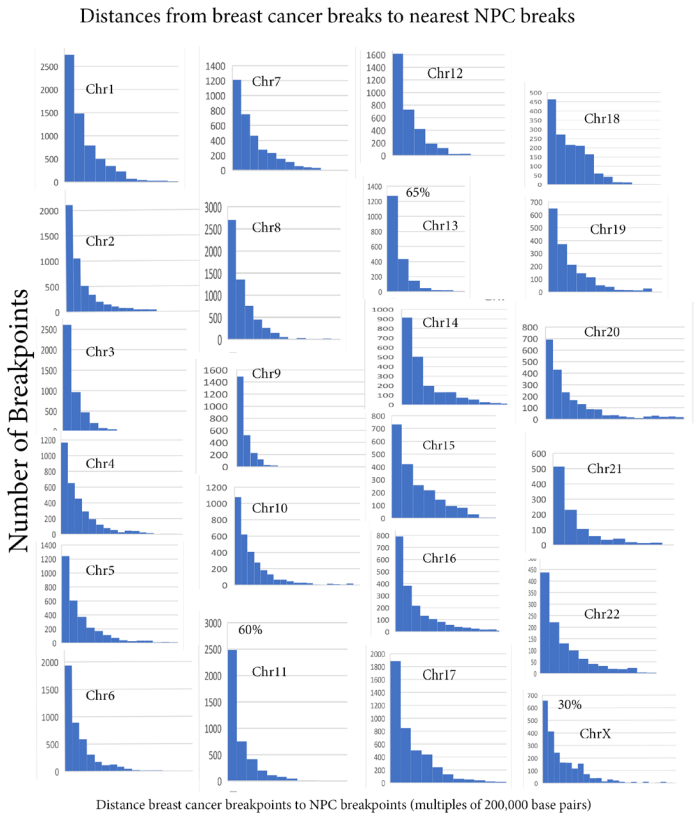


Fig. 1 Breakpoints in 139 hereditary breast cancers cluster around the same breakpoints found in 70 NPC's. On all chromosomes, breakpoints in female breast cancers are most frequently within 200k base pairs of breakpoints in NPC. Breast cancer breakpoints within 200,000 base pairs of a NPC breakpoint vary from 65% on Chr13 to 30% on ChrX. On the X axis, each bar on the graph is separated by 200,000 base pairs (the approximate length of EBV allowing for error).

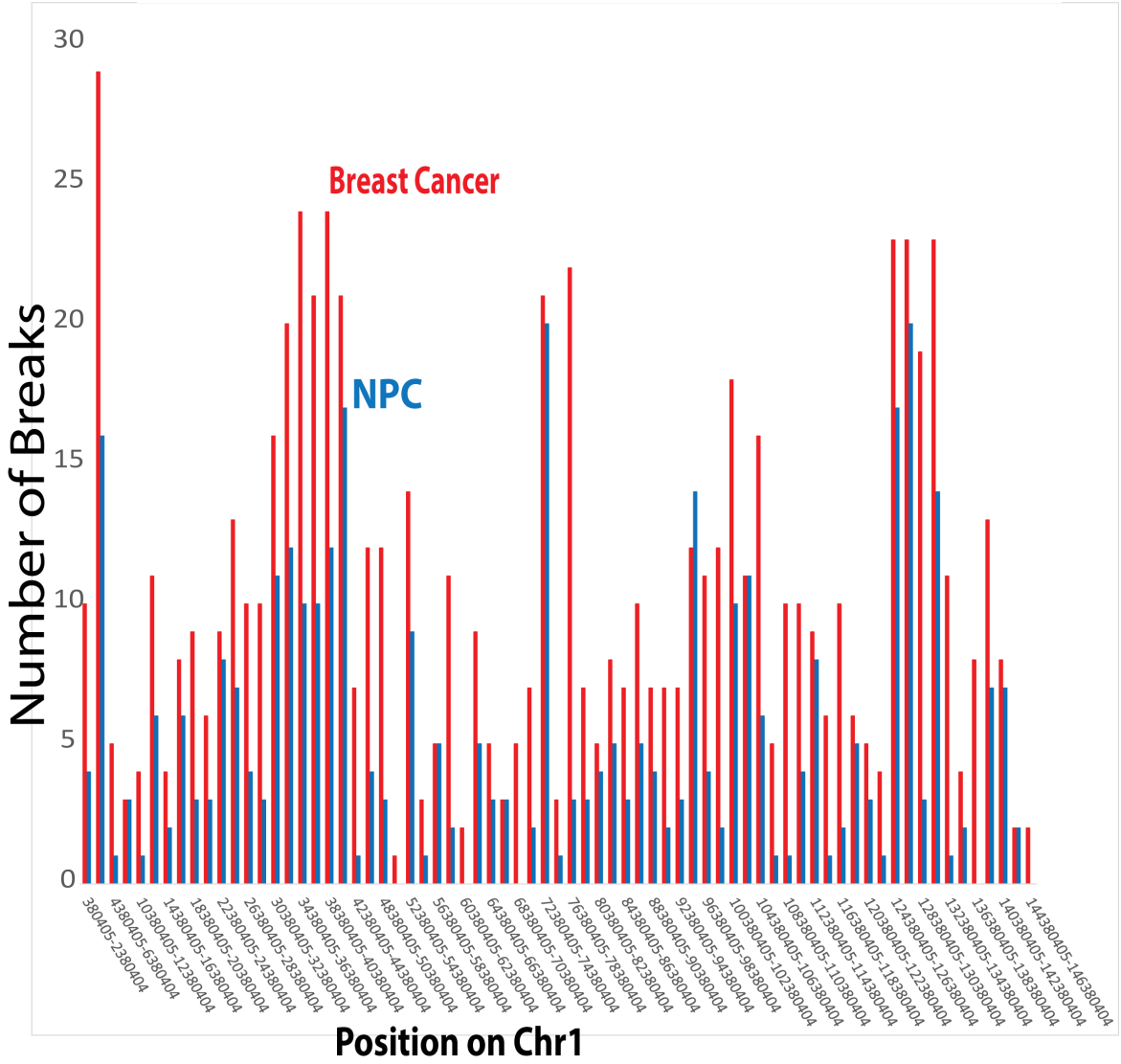
bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Based on genomes from 139 hereditary breast cancers and 70 nasopharyngeal cancers (NPC), Fig. 1 shows that every human chromosome in female breast cancers has breakpoints that cluster near those in NPC, a known EBV-mediated cancer. Peaks at the left in Fig. 1 graphs show that most breakpoints are within 200,000 base pairs of breakpoints in NPC. (200,000 is the approximate number of base pairs in EBV, allowing for some error). The exact percentage of breast cancer breaks near NPC breaks varies on different chromosomes but rises to 65% on chromosome 13.

If chromosome breakpoint positions follow probability theory, they would result from many independent events. In probability theory, the Central Limit Theorem states that the sums of independent random variables tend toward a normal distribution (i.e., a bell-shaped curve) even if the original variables themselves do not fit a normal distribution. If this logic applies to breast cancer breakpoints, they should occur at many positions, each making a small contribution. This assumption predicts a Gaussian distribution of breakpoints in the collection of breast cancers. Contrary to this prediction, statistical tests for normality found no Gaussian distribution for thousands of breakpoints in any hereditary breast cancer, sporadic breast cancer, or nasopharyngeal cancer on any chromosome. Instead, many breakpoints are shared among cancers. To test the possibility that the sharing of breakpoints is a statistical artifact, 3000 breakpoints in NPC and breast cancer were each represented by a set of random numbers between 1 and 150,000,000 (a typical chromosome length). The closest random "NPC' break point was assigned as the minimum difference between the first number below or above the random "breast cancer" breakpoint. From the two sets of 3000 random breakpoints, only 18 (0.6%) agreed within 200,000. Chromosome 1 is shown as an examples of the clustering effect of breakpoints in 139 hereditary breast cancers and 70 NPCs (Fig. 2). With few exceptions, breakpoints in the two sets of unrelated data agree well. The data is not normalized or filtered in any way.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fig 2 Breakpoints in breast cancer (red) and NPC (blue) cluster on chromosome 1. Data represents 139 breast cancers and 70 NPCs.



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

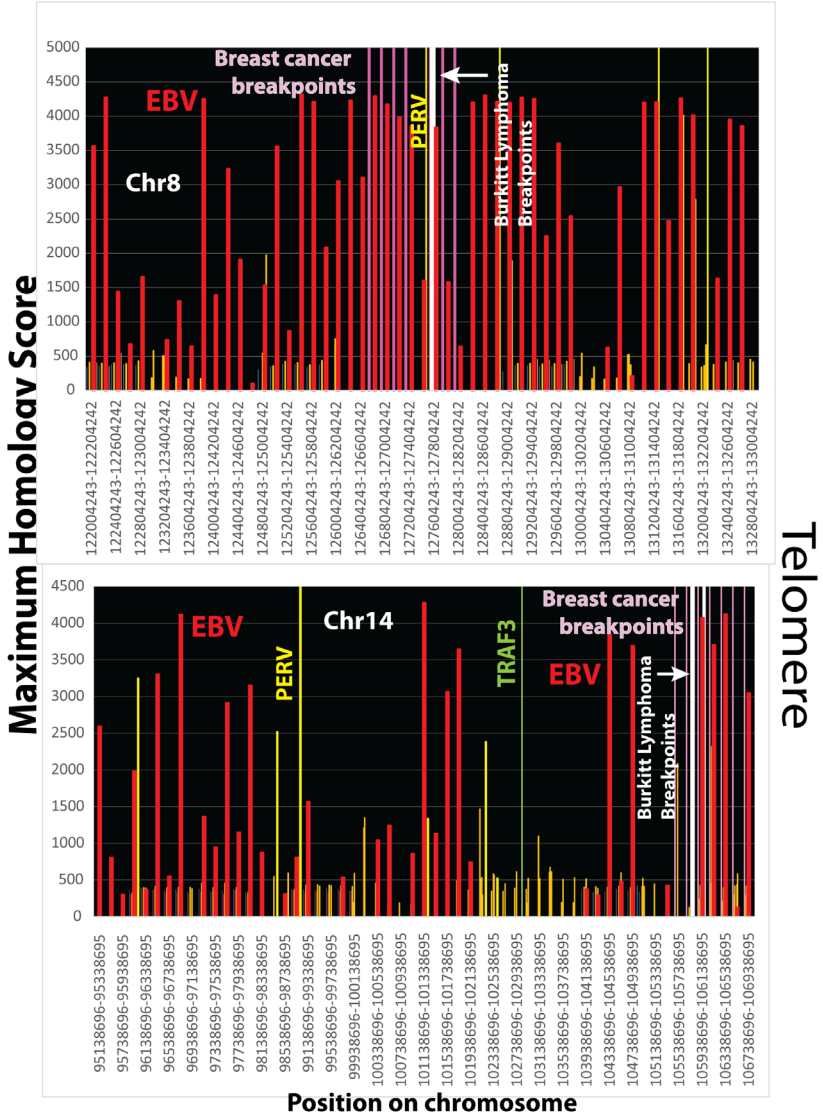


Fig. 3 Breakpoints at the MYC gene in Burkitt's lymphomas (white line), a cancer strongly associated with EBV are surrounded by DNA similar to EBV variants HKHD40 and HKNPC60 (indicated in red). Two areas of homology to porcine endogenous retrovirus (yellow lines) may also be relevant but are more distant. In the lower panel, the most common breakpoints in Burkitt's lymphoma occur at the IgM switch region near the end of the chromosome. Strong homology to EBV variants are near the DNA segment, excised during class switch recombination. Binding to EBV variants then would prevent proper excision and splicing. The TRAF3 gene is nearby and mediates the class switch from IgM to IgG. Near the telomere on the q arm of chr14, breast cancer breakpoints (pink) cluster around these regions.

Breakpoints in Burkitt's lymphoma are near EBV-like sequences. Burkitt's lymphoma is another cancer strongly linked to EBV infection. Burkitt's lymphoma has a characteristic gene rearrangement in the Myc gene on chromosome 8 (Busch et al., 2007). Fig. 3 focuses the Chromosome 8 homology data to a 10-megabyte region around the Myc gene. The

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

characteristic Myc breakpoints in Burkitt's lymphoma are surrounded by multiple DNA segments that strongly resemble EBV. In fact, 59 segments resembling viruses were within 200,000 base pairs of MYC breakpoints. Of these, 54 were from EBV-like sequences. A MYC companion breakpoint cluster in Burkitt's lymphoma often occurs in the IgM class switch region near the telomere on the q arm of chromosome 14. The class switching recombinase excises a chromosome segment between VDJ and CH1. The lower panel in Fig 3 shows that the breakpoints on the antibody H-chain gene cluster (Lopez et al., 2019) are near a region of strong homology to EBV.

Damage to genes needed for the immune system in *BRCA1* and *BRCA2* associated breast cancers and in NPC.

All the hereditary breast cancers tested have significant damage to genes needed for immune system functions, and the genes often overlap those affected by breakpoints in NPC (Fig 4). Genes affected by breakpoints in familial breast cancers were compared to 4723 genes in the innate immunity database (Breuer et al., 2013). At least 1542 innate immunity-related genes were directly affected by breakpoints in 65 familial breast cancers. The damage interferes with responses to antigens, pathogens, the ability to remove abnormal cells, and likely allows latent EBV infections to escape from control.

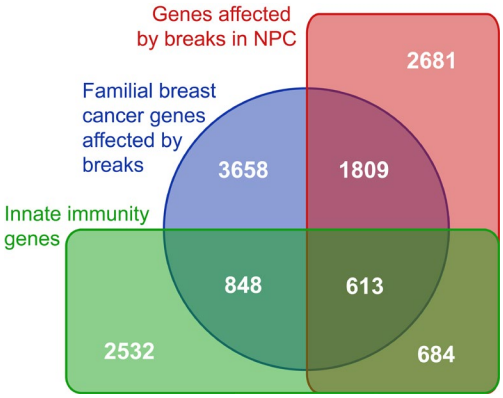


Fig. 4. Extensive overlap between genes affected by breakpoints among hereditary breast cancers, nasopharyngeal cancers and innate immunity.

Deregulation of innate immune responses may increase mutagenesis and drive multiple human cancers (Law et al., 2020). A compromised or suppressed innate immune system allows reactivation and persistent infection by populations of EBV variants. Specific genes and pathways mutated in NPC are also instrumental in driving hereditary breast cancers. Some of the breakpoints in 65 familial breast cancers directly affect these NPC gene drivers, such as NFKB1, TGF-B. For example, 19 breast cancer breakpoints directly affect NFKB, 34 affect TGF-B, and 42 affect an interferon. Conversely, just as in NPC, breast cancers constitutively activate NFKB causing an inappropriate inflammatory response (Nakshatri et al., 1997). Other herpes-mediated cancers, such as Kaposi's sarcoma also occur in the context of compromised immunity.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Breakpoints occur at human sequences that resemble herpes viruses.

Virus-human homology comparisons were conducted around thousands of human BRCA1 and BRCA2 breast cancer associated breakpoints. Long stretches of EBV variant DNA from two human gamma-herpesvirus 4 variants, HKNPC60 or HKHD40 are virtually identical to human breast cancer DNA around many inter-chromosomal breakpoints. Maximum homology scores for human DNA vs. herpes viral DNA are routinely over 4000, representing 97% identity for nearly 2500 base pairs, with E "expect" values (essentially p-values) equal to 0. The extensive homology thus represents EBV-like breakpoint signatures.

Breakpoints in hereditary breast cancers cluster around human sequences that resemble herpes virus variants. Fig. 5A shows all the viral homologies on the 145,138,636 base pairs in chromosome 8. Only a few different viruses have the strongest resemblance to human sequences. Over 11,000 regions have significant homology to EBV tumor variants. Relatively few positions (70) are similar to endogenous retroviruses, but the homology is strong.

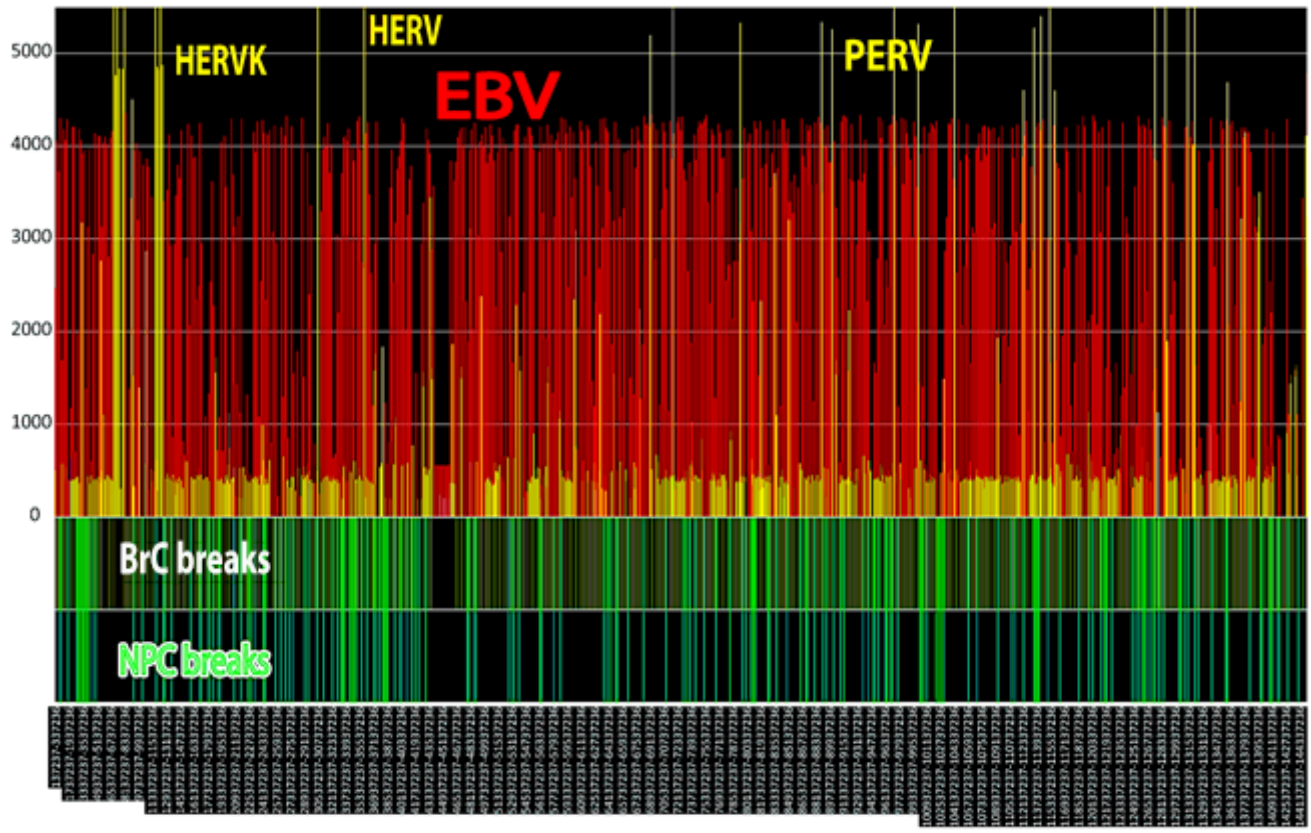


Fig. 5. Maximum homology scores to all viruses vs. position on the entire length of chromosome 8. The strongest resemblance occurs to variants of only a few viruses indicated. EBV variants clearly predominate. Break positions in hereditary breast cancers and NPC are shown.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The positions of breaks in hereditary breast cancers agree closely with the position of human sequences homologous to EBV variants and stealth viruses (likely related herpes viruses such as cytomegalovirus). On chromosome 8, nearly 6000 human EBV-like sequences are within 115,100 base pairs of a breakpoint in hereditary breast cancers, well within the number of base pairs in EBV. There are far fewer breakpoints at greater distances (Fig.5B).

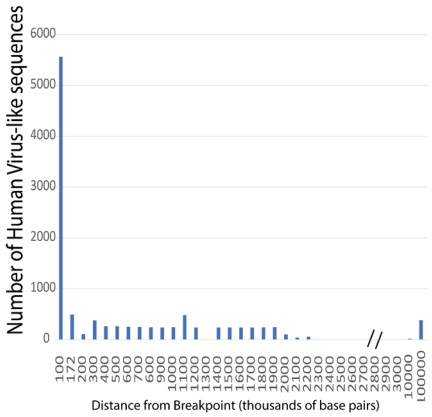


Figure 5B. Most breakpoints on chr8 are within 100kb of a virus-like sequence.

Breakpoints anywhere along the entire length of chromosome 11 were also compared to positions of homology to all viruses (Fig. 6). Chromosome 11 has 4141 inter- and intra-chromosomal breakpoints in the 139 hereditary breast cancers. Across all 135,086,622 base pairs in chromosome 11, there are 6212 matches to viral sequences having a maximum homology score above 500 within 200k base pairs of a breast cancer breakpoint. EBV variants are related to 71% of these matches. Five reiterations of a control 4141 breakpoints generated as random numbers had only 24 to 36 matches within 200,000 base pairs of a human viral-like sequence. Moreover, the 139 hereditary breast cancers had 205 matches to viral sequences within 1000 bases of a breakpoint. A random control had 0.

EBV homologies still predominate on chromosome 17 even though there are fewer matches to EBV variants than on chromosome 11. The 83,257441 base pairs on chromosome 17 have 24,206 matches with virus-like sequences with 14,859 within 200,000 base pairs of a breakpoint, while random value breakpoints have only 396. Of 2147 breakpoints with a maximum homology score above 500 within 200k base pairs of a breakpoint, only 34% (737/2147) are related to EBV. For the data from both chromosome 11 and chromosome 17, the Fisher test odds ratio that the differences from random samples were significant is 137, $p<0.0001$.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

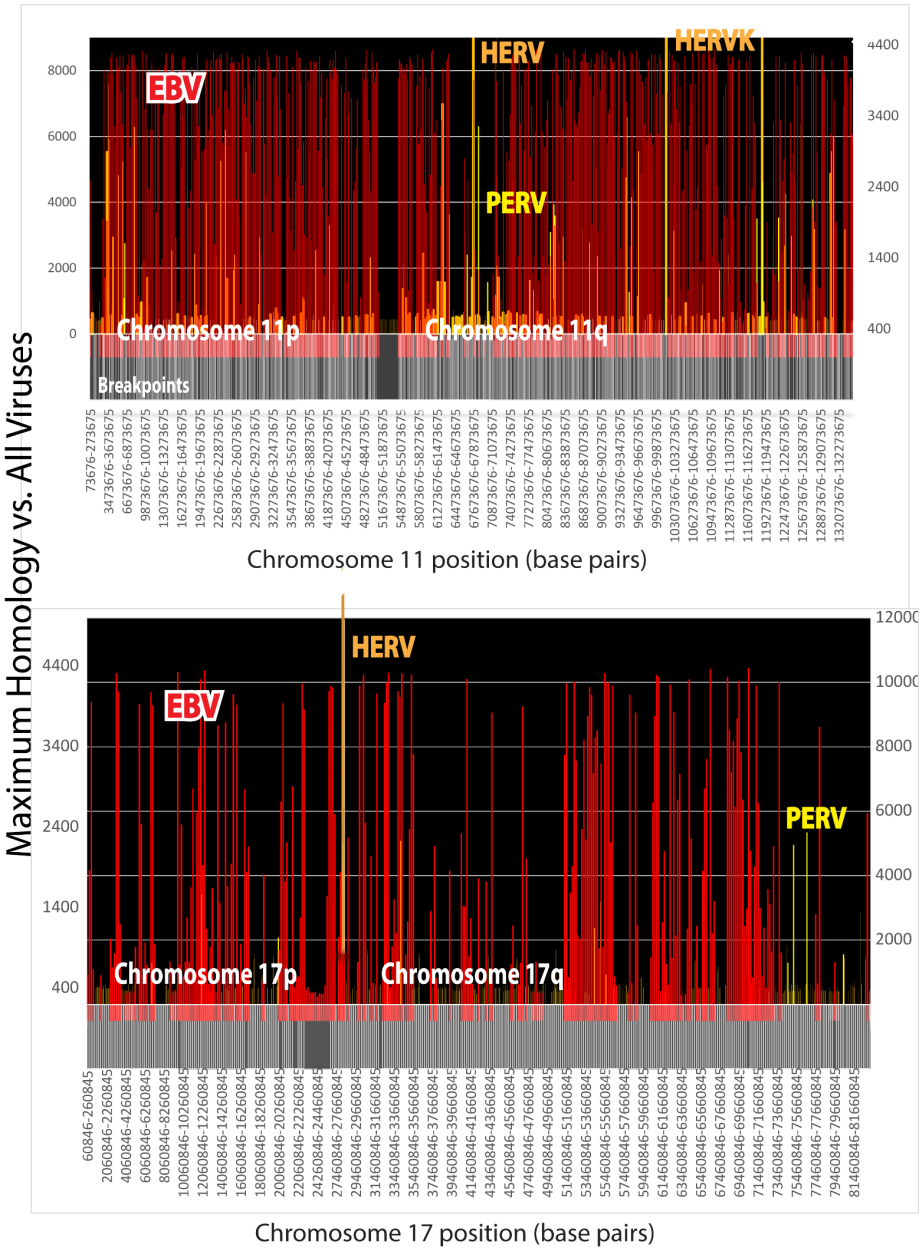


Fig. 6. Hereditary breast cancer breakpoints vs the distribution of human sequences homologous to EBV variants and endogenous retroviruses. The figure represents the entire lengths of chromosome 11 (135 million base pairs) and chromosome 17 (83.3Million base pairs).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Hereditary breast cancer breakpoint homologies to EBV are near known EBV genome anchor sites: global comparisons on two chromosomes.

Most breakpoints on segments of chromosomes 2 and 12 are near EBV genome anchor sites. On a 21 Mb section of chromosome 2, (Fig. 7a), 63% of EBV docking sites are within 200k base pairs of a breakpoint in 139 BRCA1, BRCA2 hereditary or likely hereditary breast cancers. The docking of viral DNA likely coincides with EBV anchor sites due to the similarity between the human sequence and the viruses. Start positions of human genome similarity to the EBV variant tumor viruses HKNPC60 and HKHD40 positions aligned closely to almost all 56 EBV docking sites (Fig. 7a). Some areas within this region of chromosome 2 also have significant homology to human retroviruses, porcine retroviruses, and to SARS-CoV-2 virus. A relatively low background similarity to HIV1 spreads across the region.

Clusters of breakpoints are obvious in Fig 7a (black lines at the bottom of the graph in the upper part of the figure). In every case, these clusters include EBV docking sites, regions of strong EBV homology, or areas of homology to other viruses. Chromosome 2 has 1035 breakpoints; 924 of these breakpoints were within 200,000 base pairs of homology to an EBV variant or an EBV docking site, accounting for 89% of the breaks.

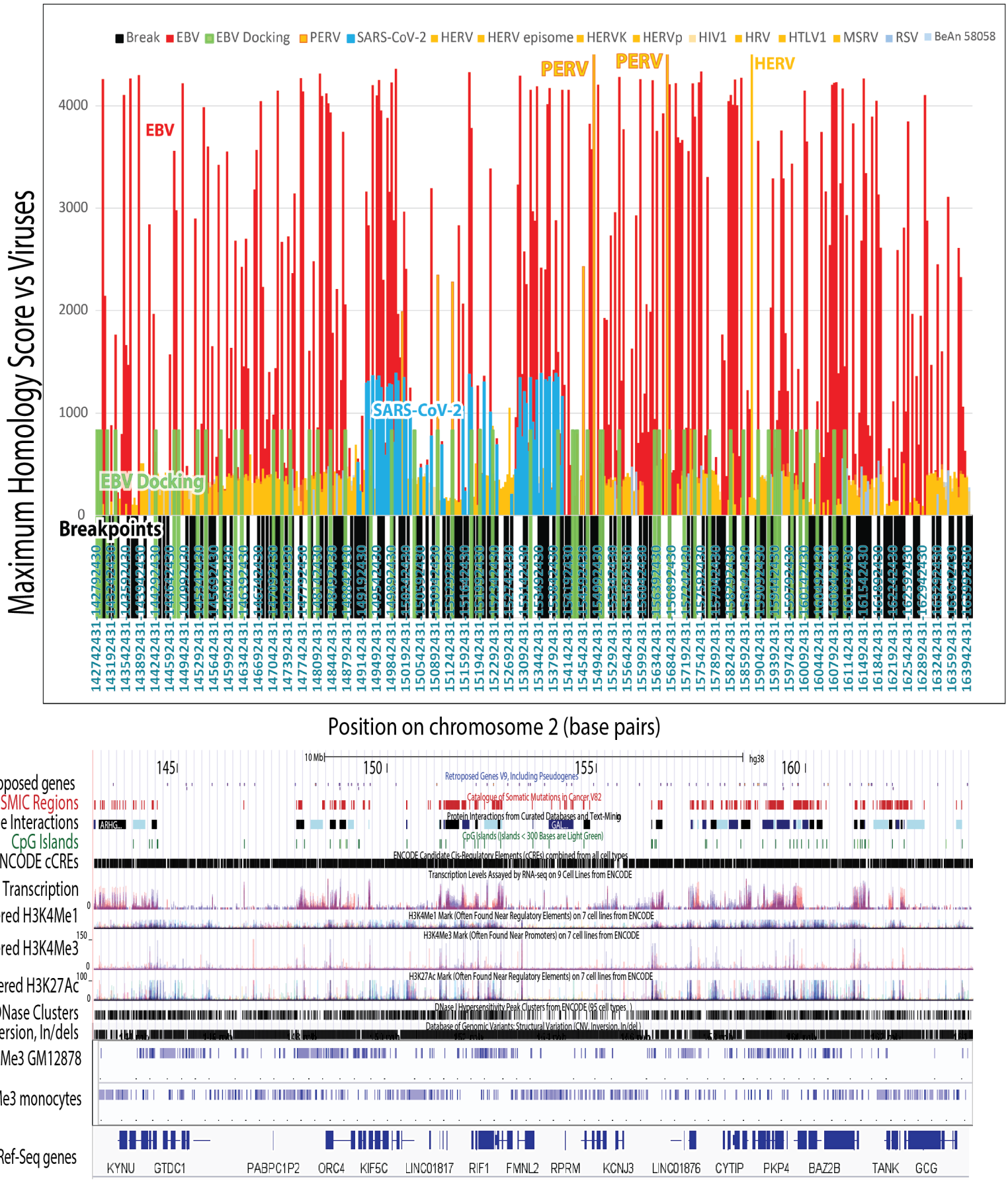
There are 180 breakpoints on a 13 Mb section of chromosome 12, with only three not near an EBV docking site. Most breaks are within 200k base pairs of a region with strong viral homology defined as a maximum homology score >=500. In all, there are 947 viral homologies, and 713 (75%) are within 200k base pairs of a break. (Fig. 7b).

The lower panels in Figs 7a and 7b support the idea that breakpoints occur in accessible regions and can have devastating effects. Breakpoints are all accessible as indicated by DNase hypersensitivity. Many breakpoints disrupt gene regulation, gene interaction, and transcription. Breakpoints all go through ENCODE candidate cis-regulatory elements (cCREs). Many breaks affect cancer-associated (COSMIC) genes. Some breakpoints disrupt the epigenetic stimulator H3K27Ac, an enhancer mark on histone packaging proteins associated with increased transcription. Both chromosome 2 and 12 sections are foci for structural variation such as CNV's, inversions, and short insertion/deletions.

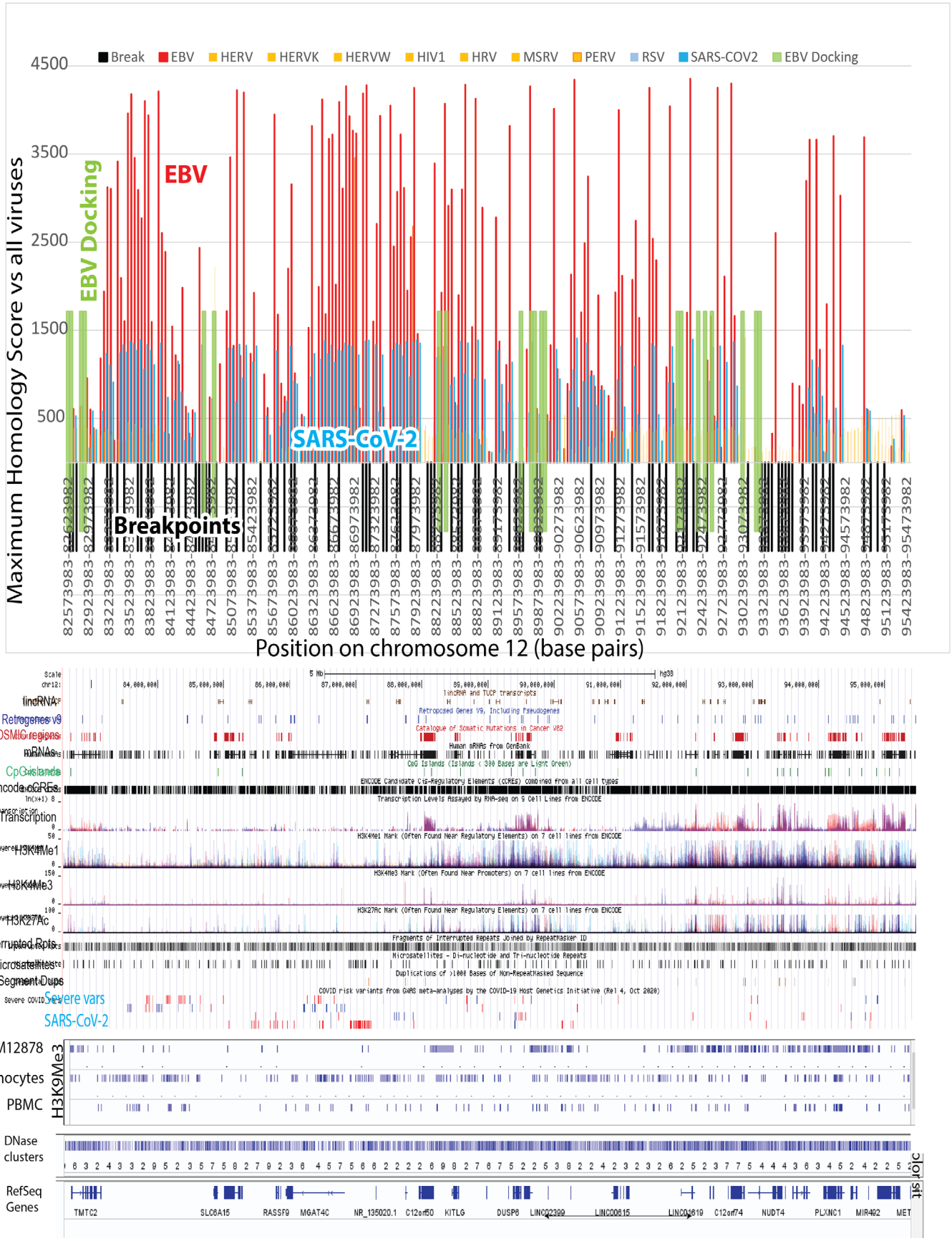
Most breast cancer breakpoints are near inhibitory epigenetic H3K9Me3 peaks in CD14+ primary monocytes (RO-01946). These markings occur around EBV genome anchor sites, where they contribute to viral latency and repress transcription (Kim *et al.*, 2020). Both regions on chromosomes 2 and 12 are rich in these sites (Figs. 7a and 7b).

Multiple breakpoints on either chromosome 2 or 12 disrupt reference genes. Human reference sequence genes in the chromosome 2 region include *KYNU, GTDC1, ACVR2A, KIF5C, STAM2, KCNJ3, ERMN, PKP4, BAZ2B, TANK,* and *DPP4* (Fig. 7a, bottom). Breast cancer breaks near at least some of these genes interrupt functions essential for immunity and preventing cancer. For example, *KYNU* mediates the response to IFN-gamma. *TANK* is necessary for *NFKB* activation in the innate immune system. *DPP4* is essential for preventing viral entry into cells. Reference gene functions in the breakpoint region of chromosome 12 (Fig. 7b, bottom) include vesicle trafficking (*RASSF9*), endocytosis (*EEA1*), blood cell formation (*KITLG*), interferon response control (*SOCS2*), and nerve cell patterning (*NR2C1*).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Fig. 7(a), Breakpoints on a 21 million base pair segment in human chromosome 2 in 139 BRCA1, BRCA2, and presumptive BRCA-associated breast cancers. EBV anchor locations (green lines) and breast cancer breakpoints (black lines) are shown to allow comparisons. In the top panel, the axis at the left represents maximum homology to all viruses. The lower section shows characteristics of the human genome aligned with selected regions from the UCSC genome browser. In Fig 7b, breakpoints from 139 hereditary or likely hereditary breast cancers on a 13 million base pair segment of chromosome 12 are aligned with homology to all viruses and to selected values from the UCSC genome browser.

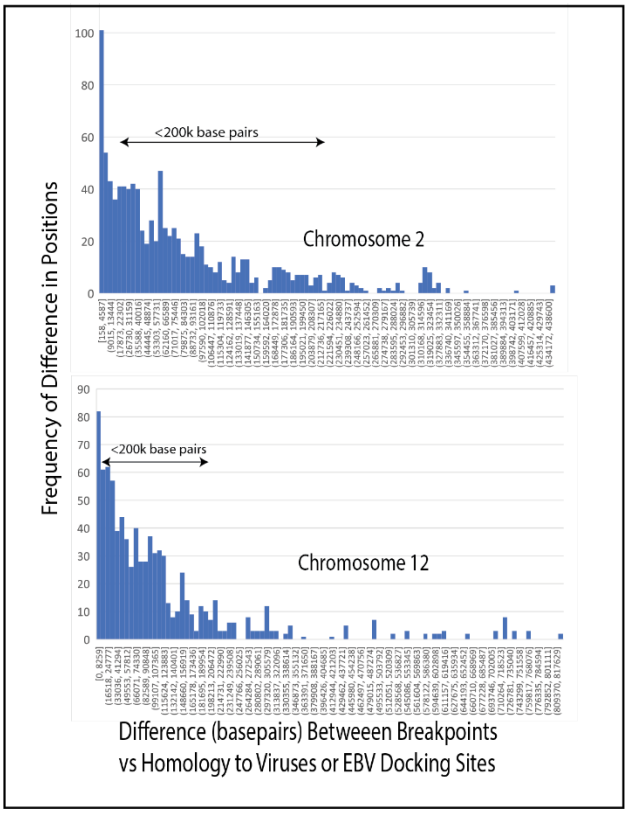


Fig. 7c shows how often breakpoints in 139 hereditary or likely hereditary breast cancers on chromosome 2 and chromosome 12 are near areas of homology to viruses or EBV docking sites.

Endogenous retrovirus-like DNA sequences are also near breakpoints but distribute differently on chromosome 2 vs. chromosome 12. (Figs. 7a vs. 7b). On chromosome 2, porcine endogenous retrovirus (PERV) (Denner, 2017), human endogenous retrovirus (HERV) sequences (Prusty et al., 2008), and a pseudogene (pHERV) also have significant homology to human DNA. The PERV-like sequence lies within a retroposed area on chromosome 2 within 28 Kbps 5' and 80 Kbps 3' of EBV sequences. In contrast, significant retroviral sequence homologies are outside the breakpoint-rich stretch on chromosome 12 (Fig. 7b).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Identified EBV genome anchors are near known genes that match breast cancer breakpoints.

Fig. 8 uses independent data to establish consistent relationships of breast cancer breaks to EBV genome anchor sites, precisely identified at disparate chromosome or gene locations (Lu *et al.*, 2010). A primary EBV genome binding site on chromosome 11 (Lu *et al.*, 2010) closely matches a few breast cancer breakpoints. Another known binding site on chr1 near the CDC7 gene also corresponds to multiple breast cancer breaks.

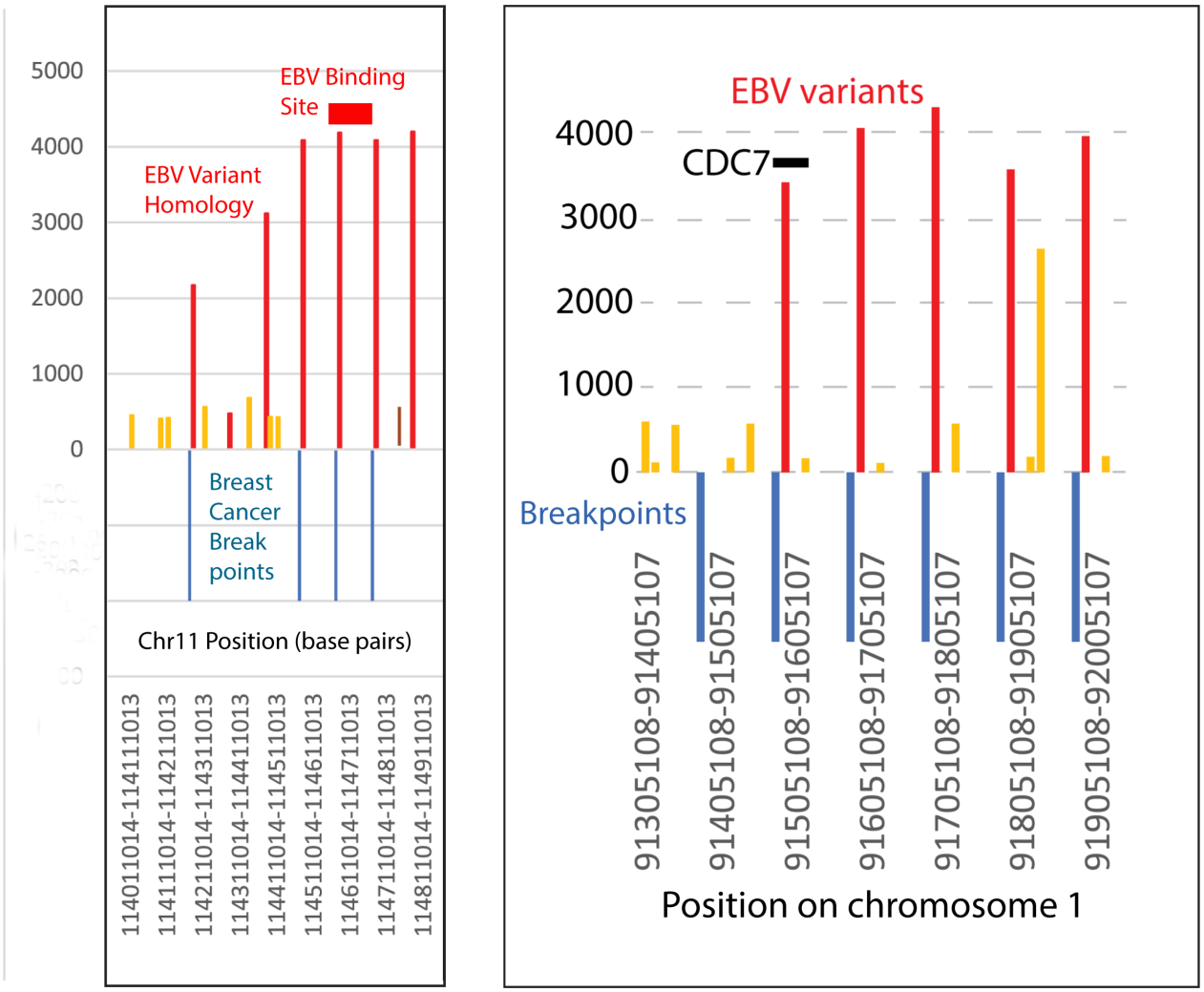


Fig. 8. Maximum homology to human DNA for all viruses (y-axis) is plotted for known EBV genome anchor sites vs. breast cancer breakpoints and gene coordinates. At the left, a primary EBV binding site are close to several breast cancer breaks and human DNA that resembles EBV variants.. At right, EBV variant homologies near CDC7 are similarly near a known EBV binding site and breast cancer breakpoints. Red bars indicate human sequences similar to EBV and orange bars are human sequences similar to endogenous retroviruses. Other known EBV anchor sites gave similar results.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

BRCA1 or BRCA2 mutations by themselves are not sufficient to cause chromosome breaks

The possibility was tested that BRCA1 and BRCA2 mutations are sufficient to cause chromosome breaks without contributions from herpes variants or other viruses. Breakpoints in breast cancers from 74 women over age 70 with no known hereditary BRCA1 and BRCA2 mutations (Nik-Zainal *et al.*, 2016) were compared to hereditary breast cancers. The female sporadic breast cancer patients are older than hereditary breast cancer patients, so mutations have had more time to accumulate, i.e., some base substitution signatures positively correlate with age (Alexandrov et al., 2020). Yet, inter-chromosomal translocation breakpoints also occur in sporadic cancers with normal BRCA1 and BRCA2 genes (Fig. 9). Although there are significant differences in breakpoint distributions (Fig. 9), many hereditary and sporadic breast cancer breakpoints cluster at the same chromosome locations. Important potential confounders include promoter hypermethylation which provides an alternate method of inactivating BRCA1 and BRCA2 genes. BRCA1 and BRCA2 methylation is unusual in sporadic cancers because average methylation scores in 1538 sporadic breast cancers (Batra *et al.*, 2021) were 0.050 and 0.004 for BRCA1 and BRCA2 promoters, respectively. Triple-negative breast cancer may be another significant confounder because up to 58.6% of 237 patients had a significant marker predictive of BRCA1/BRCA2 deficiency (Staaf *et al.*, 2019). However, triple-negative breast cancers comprised only 9% of breast cancers in the cohort.

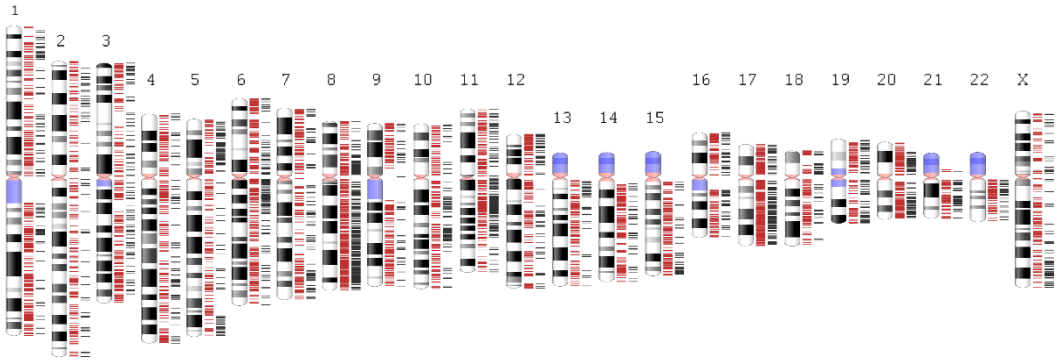


Figure 9. Inter-chromosome translocation break positions in 74 mutation-associated or likely mutation associated female breast cancers (red) vs. 74 likely sporadic female breast cancers (black).

Breakpoints in sporadic breast cancers on every chromosome are similar to breakpoints in NPC, a known EBV-related cancer.

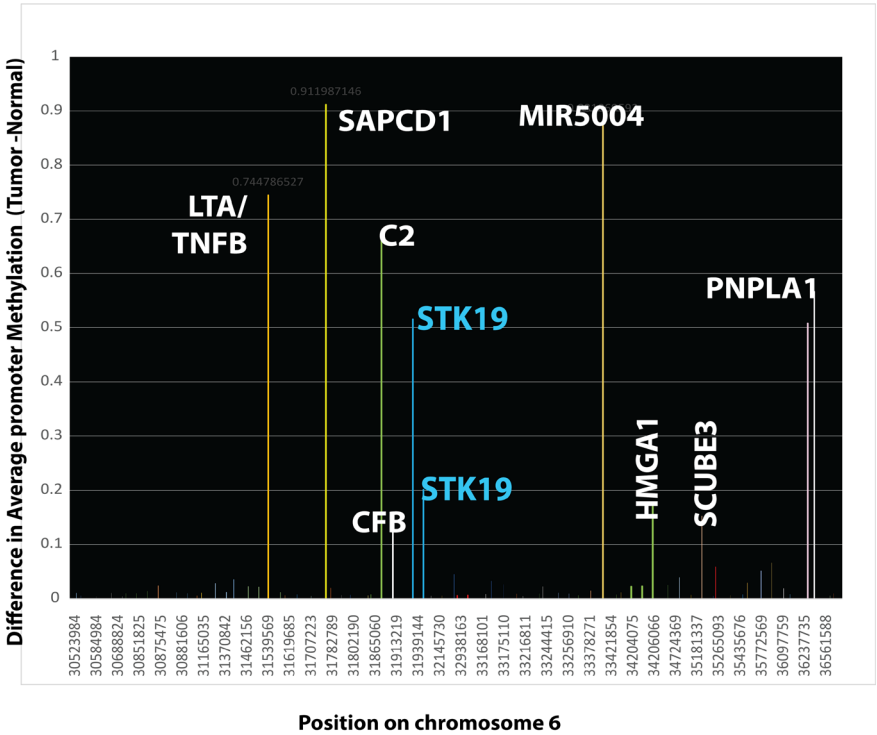
The breakage-fusion-bridge cycle is a catastrophe sometimes related to telomeric dysfunction or a break near the end of a chromosome. During the generation of tumors, the cycle causes clustering of chromosome breakpoints, chromothripsis (Leibowitz et al., 2015; Umbreit *et al.*, 2020) and kataegis while telomere dysfunction promotes end-to-end fusions. For 74 breast cancers in women over age 70, sporadic cancer breakpoint positions were compared to breakpoints in 70 nasopharyngeal cancers as a model for EBV-mediated breakages. These analyses establish a robust and reliable association between sporadic breast cancer breaks and breakpoints in NPC. This association occurs without BRCA1 and BRCA2 gene mutations. The data (not shown) generally resembles the results for hereditary cancers (Fig.1). Generally

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

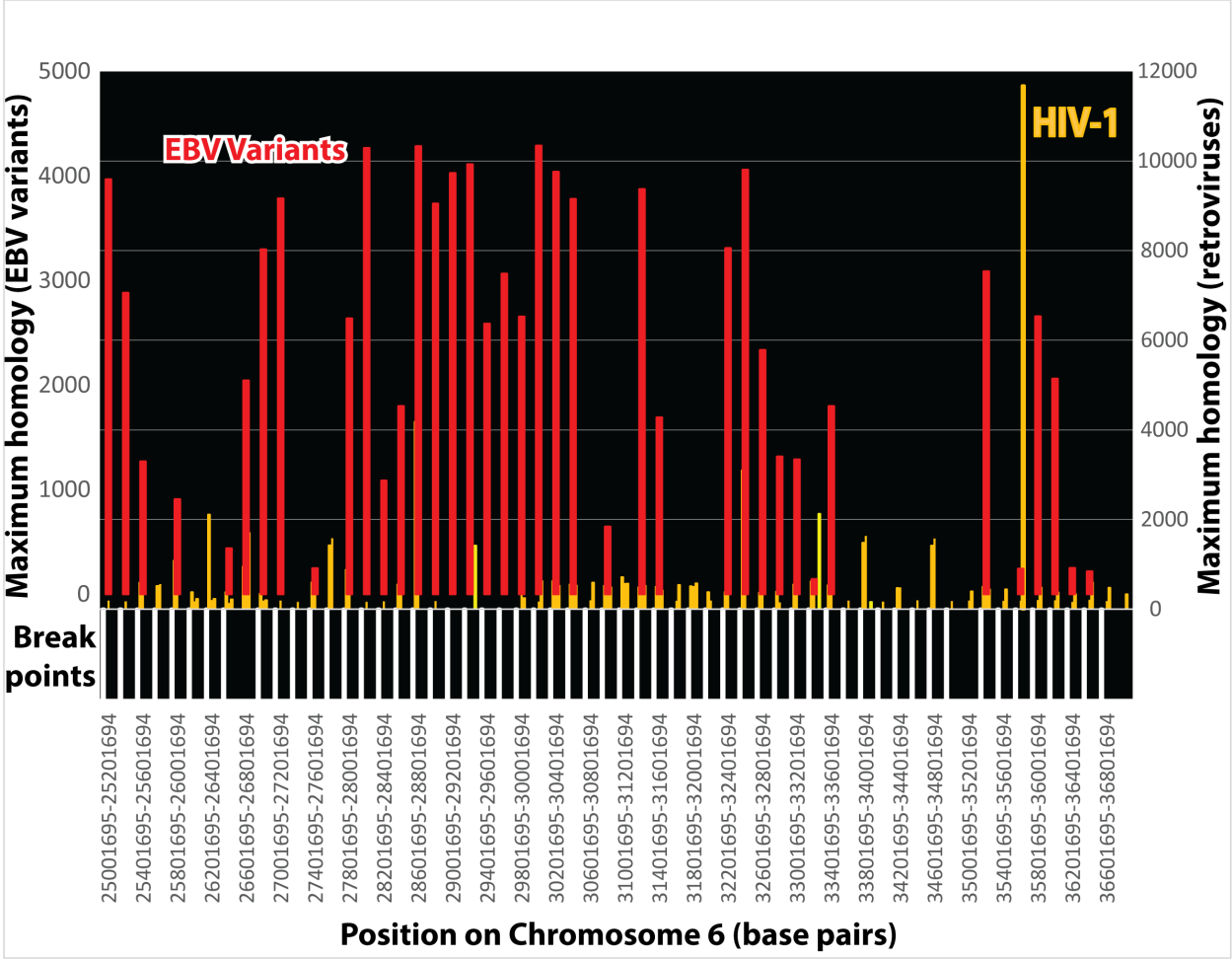
breakpoints are less common in sporadic cancers than in hereditary cancers, they are still related to those in NPC For example, sporadic breast cancer and NPC show a close relationship to 78% of breakpoints on chromosome 11.

A potential EBV signature is present in sporadic breast cancers.

Although EBV-positive NPC and gastric cancers have distinctive patterns of genes with DNA hypermethylation, some frequently DNA hypermethylated genes are shared. Chromosome 6p21.3 is a potential EBV infection signature, because the 6p21.3 region (Chr6:30,500,001-36,600,000) is hypermethylated in EBV-positive NPC and gastric cancer (Scott, 2017). To determine if this potential marker is also hypermethylated in breast cancers, methylation data from 1538 breast cancers (Batra *et al.*, 2021) was tested. Fig. 10 shows that this marker region in breast cancers has significant differences in promoter methylation vs. normal controls. Gene promoters on 6p21.3 inhibited by hypermethylation primarily control complement function, a system that integrates innate and adaptive immune responses against challenges from pathogens and abnormal cells. Moreover, most of the DNA breaks in this region are close to regions of homology to EBV sequences.



bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.



Herpes virus sequences are not an artifact

The EBV tumor viruses (HKHD40 and HKNPC60) are typical of many other herpesvirus isolates, with some haplotypes conferring a high NPC risk (Xu *et al.*, 2019a). About 100 other gamma herpes viral variants strongly matched HKHD40 and HKNPC60 in regions with enough data to make comparisons possible. HKNPC60 is 99% identical to the EBV reference sequence at bases 1-7500 and 95% identical at bases 1,200,000-1,405,000. HKHD40 is 99% and 98% identity for comparisons to the same regions.

Fragile site breaks do not account for breast cancer breakpoints

Lu et al. found 4785 EBNA1 binding sites with over 50% overlapping potential fragile sites as a repetitive sequence element (Lu et al., 2010). Kim et al. reported that EBNA1 anchor sites have A-T rich flanking sequences, with runs of consecutive A-T bases (Kim et al., 2020).

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Based on the fragile site database, chromosome 1 contains 658 fragile site genes, the most of any chromosome (Kumar et al., 2019). Although some fragile sites align with breast cancer breaks on chromosome 1, large numbers of breaks on chromosomes 4, 12 (Fig. 11) and most other chromosomes are inconsistent with the fragile site database. On all the chromosomes tested, there were many more breakpoints than fragile sites. Breast cancer breaks do not consistently occur near common fragile sites (Fig. 11). Some hereditary breast cancer breakpoints were tested for repeats likely to generate fragile sites because the repeats are difficult to replicate. This test did not find such sequences (supplementary Table S1). In contrast, many interchromosomal breaks are close to human EBV-like sequences. Sites of replication errors in even one cell can be a sudden catastrophe that cascades into further breaks, destabilizing the entire genome (Umbreit et al., 2020). This cascade is especially likely in hereditary breast cancers with their deficits in homologous recombination repair.

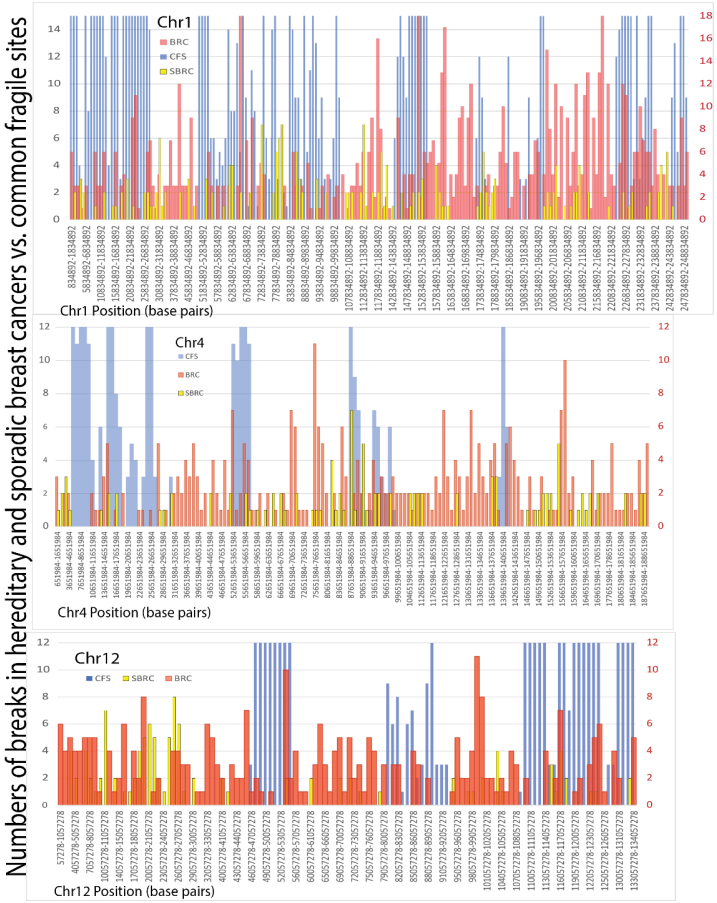


Fig 11. Histograms showing the relative positions of breakpoints in BRCA1 or BRCA2 associated breast cancers on chromosome 1, 4 and 12 vs the positions of common fragile site sequences (blue). Each chromosome was divided into 200 bins with values shown on the horizontal axis. The vertical axis is the number of times breakpoints in hereditary (red) or sporadic (yellow) breast cancer or CFS values fell into each bin.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Discussion

Most human cancer viruses merely initiate or promote cancer and do not cause cancer by themselves. The present work shows human sequences related to tumor viral variants cluster around chromosome breaks in breast cancers. Multiple independent lines of evidence support this view and are summarized below (1-12).

1. Breakpoints in both hereditary and sporadic breast cancers match breakpoints in NPC, a epithelial cell cancer mediated by EBV variants. This matching is true for every chromosome in females.
2. Breast cancer breakpoints and EBV like sequences on chromosome 14 cluster around the IgM class switch region characteristic of Burkitt's lymphoma. Similarly MYC region breakpoints in Burkitt's lymphoma are surrounded by EBV like sequences and a cluster of breast cancer breakpoints.
3. A potential EBV methylation signature shared with known EBV cancers is far more abundant in 1538 breast cancers than in normal controls.
4. Breast cancer breakpoints are consistently near binding sites for EBV.
5. Every human chromosome in female breast cancers has breakpoints that are close to EBV-like human DNA. EBV can also activate endogenous retroviral sequences, which also occur near some breakpoints.
6. The association of breast cancer breaks and EBV variants does not depend on the presence of BRCA1 or BRCA2 gene mutations. However breast cancer breaks are more abundant in BRCA1 and BRCA2 mutation carriers than in non-carriers.
7. Sets of random breakpoints do not collect around similar values, like breakpoints in breast cancer and NPC.
8. The association of EBV variant sequences with chromosome breakpoints does not require the continuing presence of active viruses anywhere within the resulting tumor. One breakpoint in a single cell can generate further breaks during cell division and destabilize the entire human genome. Clusters of mutation also occur after illicit break repair.
9. Deficient immune responses in breast cancer tissue make it unduly susceptible to reactivation of exogenous EBV-like infections. The tissue also becomes unable to remove abnormal cells Some breakpoints directly disrupt essential immune response genes.
10. Fragile site sequences are not sufficient to account for breast cancer breaks
11. Tumor variants HKHD40 and HKNPC60 are herpesviruses closely related to known tumor virus populations KHSV and EBV.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12. EBV infection is nearly universal and can cause genomic instability well before breast cancer occurs.
13. EBV can also activate endogenous retroviral sequences, which also occur near some breakpoints.
14. Epidemiologic data shows active EBV infection increases breast cancer risk by about five-fold over controls. This estimate is probably conservative because it does not consider viral variants.

Cancer drug therapy focuses on identifying cancer driver mutations and then using drugs that will target them. The drugs are initially effective but then quit working. This is because the drugs target by-products of an occult, underlying process, but the process continues even if cells with the initial driver mutations are removed. Fig. 12 summarizes a proposed model that explains why cancer relapses after initial remission. Chromosomes from populations of EBV variants infecting an individual bind to chromosome breaks in human DNA at sequences matching variants. Under the best conditions, the break is quickly repaired by rejoining the two fragments. However the repetitive virus-like human sequences fuse to inappropriate chromosome fragments using bound, ubiquitous EBV sequences as a guide. EBV DNA gets packaged around chromatin, as if it were a small human chromosome. The viral chromosome moves around through the nuclear environment of human chromosomes depending on the state of viral activation. EBV in its latent state binds human repressive heterochromatin, but viral reactivation moves it to active euchromatin (Moquin et al., 2018). Plentiful EBV binding sites exist as homologous repetitive human DNA sequences that are nearly 2500 base pairs long. DNA near breaks is likely to be accessible because helicases and other enzymes expose active chromatin (Aleksandrov et al., 2020). Exposure is necessary so that repair processes can access broken DNA. EBV variants can also cause breaks more directly by producing nucleases (Wu et al., 2010), promoting telomere dysfunction and replication stress (Hafez and Luftig, 2017).

A broken chromosomes is deficient in telomeres, so a second chromosome fragment can link to the first one, but the product now has two centromeres. During anaphase in cell division, the two centromeres try to separate, form a bridge and the chromosome breaks again (McClintock, 1941). The bridge region does not replicate normally during mitosis so chromothripsis and kataegis become integral in the breakage-fusion-bridge mechanism (Umbreit *et al.*, 2020). Translocations can generate a burst of localized somatic mutations through the actions of APOBEC ("kataegis") (Nik-Zainal and Morganella, 2017). APOBEC is typically a response to inactivate viral infections and effects related to APOBEC3 probably occur in EBV-induced carcinogenesis (Bobrovnitchaia et al., 2020; Law *et al.*, 2020). The process repeats during cell division, quickly increasing the numbers of abnormal chromosome products and driver mutations.

BRCA1 and BRCA2 mutations increase the numbers of breaks, but breakage-fusion-bridge cycles also occur in sporadic breast cancer chromosomes, albeit less frequently. Fragments of chromosomes such as those generated during chromothripsis can also join different chromosomes if their ends are not protected by telomeres. Support for this model comes from the observation that replication is an active genomic process that encourages breaks (Cheloshkina and Poptsova, 2021). In addition to gene defects in BRCA1 and BRCA2, other inherited gene defects

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

increase susceptibility to EBV infection and EBV-driven diseases.. These inherited forms are associated with mutations in SH2D1A, ITK, MAGT1, CTPS1, CD27, CD70, CORO1A, and RASGRP1 (Latour and Winter, 2018). Over 50% of patients with one of these defects experience EBV-driven lymphoproliferative disease including Hodgkin and non-Hodgkin lymphomas. Severe viral infections with other herpes viruses (CMV, HSV, HHV-6) are also common.

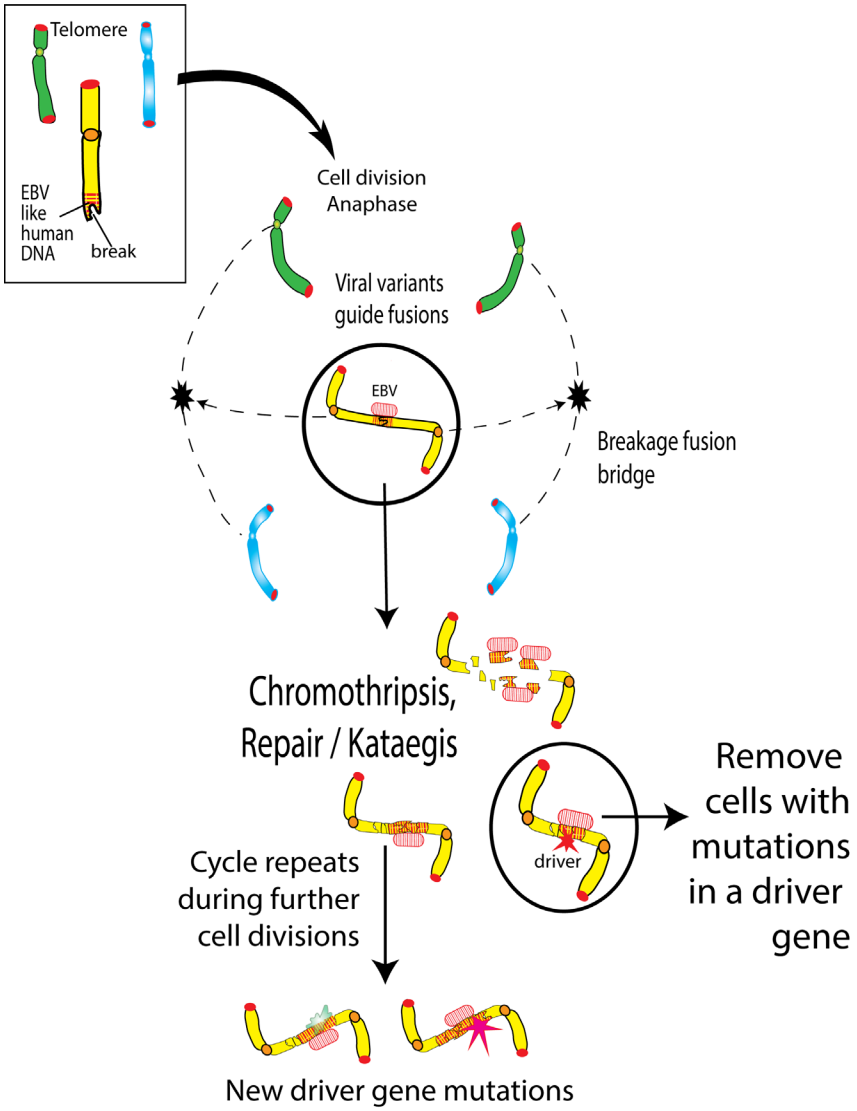


Fig 12. *Human sequences around cancer breakpoints match viral variants. Viruses can guide fusions and even cause breaks. Removing cells with driver mutations cannot stop breakage fusion bridge cycles so initially effective cancer drugs stop working.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

The sheer numbers of strong EBV-like sequences in human DNA are surprising and rival retroviruses homologies. This similarity reflects the constant interplay between herpes viral DNA and human DNA during evolution. Some of these interactions are already recognized as causing severe disease. For example, a related herpes virus HHV-6A/6B integrates into telomeres on every chromosome. The viruses affect about 1% of humans (Tweedy et al., 2016), causing severe disease on reactivation. Cytomegalovirus, another related herpes virus, causes birth defects. In many cases, PERV or other endogenous retroviral sequences also contribute to breast cancer breaks. Significant homology to SARS-CoV-2 also exists (Figs 7a and 7b) on Chromosomes 2 and 12. Host factors are the primary determinant of the severity of SARS-CoV-2 infection (Zhang et al., 2020). These areas of homology together with deregulated immune responses, may add genomic host factors that influence the severity of CoV-2 infection.

Multiple components of immune defenses mutate in hereditary breast cancer genomes. The mutations affect processes such as cytokine production, autophagy, etc. These functions depend on many genes dispersed throughout the genome, so any cancer needs only to damage one gene to cripple an immune function. Each breast cancer genome has a different set of these mutations, with the same gene only occasionally damaged (Friedenson, 2013; 2015). Damage affecting the nervous system is also universal. Some herpes viruses establish occult infection within the central nervous system even after other sites become virus-free (Bhela et al., 2014). The immune system and the nervous system communicate extensively. Neurotransmitters from parasympathetic and sympathetic neurons control immune activity. For example, TLR3 stimulation by viral infection triggers cytokine production from neurons that promotes immune responses. Injury, autoimmune conditions, hypoxia, and neurodegeneration activate immune cells in the central nervous system (Kioussis and Pachnis, 2009)

The results of the present work are potentially actionable. The current evidence adds support for cancer prevention by a childhood herpes vaccine and cancer treatment by EBV antivirals. The prospects for producing an EBV vaccine are promising, but the most appropriate targets are still not settled. Some immunotherapy strategies rely on augmenting the immune response, but this approach may need modification because mutations create additional holes in the immune response. Retroviruses and retrotransposons (Helman et al., 2014) may also participate in breast cancer breaks. Participation from porcine endogenous retroviruses is actionable by thoroughly cooking pork products. However, despite assertions that xenotransplantation with pig cells is safe, it is concerning that up to 6500 bps in human chromosome 11 are virtually identical to pig DNA (Fig. 7a).

Sampling the population to represent the breadth of all somatic and hereditary breast cancers is a significant problem and limits manystudies. There is no assurance that even large numbers of breast cancers are an adequate representation because they are not a random sample from all breast cancers (Friedenson, 2009). The data used here comes from 560 breast cancer genome sequences, familial cancer data from 78 patients, methylation data from 1538 breast cancers vs 244 controls, and 243 triple negative breast cancers (Batra *et al.*, 2021; Nik-Zainal *et al.*, 2016; Nones *et al.*, 2019; Staaf *et al.*, 2019).

References

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Aleksandrov, R., Hristova, R., Stoynov, S., and Gospodinov, A. (2020). The Chromatin Response to Double-Strand DNA Breaks and Their Repair. Cells *9*. 10.3390/cells9081853.
Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al. (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94-101. 10.1038/s41586-020-1943-3.
Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecular biology *215*, 403-410. 10.1016/S0022-2836(05)80360-2.
Ayee, R., Ofori, M.E.O., Wright, E., and Quaye, O. (2020). Epstein Barr Virus Associated Lymphomas and Epithelia Cancers in Humans. Journal of Cancer *11*, 1737-1750. 10.7150/jca.37282.
Balfour, H.H., Jr., Sifakis, F., Sliman, J.A., Knight, J.A., Schmeling, D.O., and Thomas, W. (2013). Age-specific prevalence of Epstein-Barr virus infection among individuals aged 6-19 years in the United States and factors affecting its acquisition. The Journal of infectious diseases *208*, 1286-1293. 10.1093/infdis/jit321.
Batra, R.N., Lifshitz, A., Vidakovic, A.T., Chin, S.F., Sati-Batra, A., Sammut, S.J., Provenzano, E., Ali, H.R., Dariush, A., Bruna, A., et al. (2021). DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. Nat Commun *12*, 5406. 10.1038/s41467-021-25661-w.
Bhela, S., Mulik, S., Reddy, P.B., Richardson, R.L., Gimenez, F., Rajasagi, N.K., Veiga-Parga, T., Osmand, A.P., and Rouse, B.T. (2014). Critical role of microRNA-155 in herpes simplex encephalitis. Journal of immunology *192*, 2734-2743. 10.4049/jimmunol.1302326.
Bobrovnitchaia, I., Valieris, R., Drummond, R.D., Lima, J.P., Freitas, H.C., Bartelli, T.F., de Amorim, M.G., Nunes, D.N., Dias-Neto, E., and da Silva, I.T. (2020). APOBEC-mediated DNA alterations: A possible new mechanism of carcinogenesis in EBV-positive gastric cancer. International journal of cancer. Journal international du cancer *146*, 181-191. 10.1002/ijc.32411.
Breuer, K., Foroushani, A.K., Laird, M.R., Chen, C., Sribnaia, A., Lo, R., Winsor, G.L., Hancock, R.E., Brinkman, F.S., and Lynn, D.J. (2013). InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. Nucleic acids research *41*, D1228-1233. 10.1093/nar/gks1147.
Bruce, J.P., To, K.F., Lui, V.W.Y., Chung, G.T.Y., Chan, Y.Y., Tsang, C.M., Yip, K.Y., Ma, B.B.Y., Woo, J.K.S., Hui, E.P., et al. (2021). Whole-genome profiling of nasopharyngeal carcinoma reveals viral-host co-operation in inflammatory NF-kappaB activation and immune escape. Nat Commun *12*, 4193. 10.1038/s41467-021-24348-6.
Busch, K., Keller, T., Fuchs, U., Yeh, R.F., Harbott, J., Klose, I., Wiemels, J., Novosel, A., Reiter, A., and Borkhardt, A. (2007). Identification of two distinct MYC breakpoint clusters and their association with various IGH breakpoint regions in the t(8;14) translocations in sporadic Burkitt-lymphoma. Leukemia *21*, 1739-1751. 10.1038/sj.leu.2404753.
Charoentong, P., Finotello, F., Angelova, M., Mayer, C., Efremova, M., Rieder, D., Hackl, H., and Trajanoski, Z. (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. Cell reports *18*, 248-262. 10.1016/j.celrep.2016.12.019.
Cheloshkina, K., and Poptsova, M. (2021). Comprehensive analysis of cancer breakpoints reveals signatures of genetic and epigenetic contribution to cancer genome rearrangements. PLoS Comput Biol *17*, e1008749. 10.1371/journal.pcbi.1008749.
Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423. 10.1093/bioinformatics/btp163.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Cuceu, C., Hempel, W.M., Sabatier, L., Bosq, J., Carde, P., and M'Kacher, R. (2018). Chromosomal Instability in Hodgkin Lymphoma: An In-Depth Review and Perspectives. Cancers (Basel) *10*. 10.3390/cancers10040091.

Denner, J. (2017). The porcine virome and xenotransplantation. Virology journal *14*, 171. 10.1186/s12985-017-0836-z.

Eliassen, E., Lum, E., Pritchett, J., Ongradi, J., Krueger, G., Crawford, J.R., Phan, T.L., Ablashi, D., and Hudnall, S.D. (2018). Human Herpesvirus 6 and Malignancy: A Review. Front Oncol *8*, 512. 10.3389/fonc.2018.00512.

Fina, F., Romain, S., Ouafik, L., Palmari, J., Ben Ayed, F., Benharkat, S., Bonnier, P., Spyratos, F., Foekens, J.A., Rose, C., et al. (2001). Frequency and genome load of Epstein-Barr virus in 509 breast cancers from different geographical areas. British journal of cancer *84*, 783-790. 10.1054/bjoc.2000.1672.

Friedenson, B. (2009). Dewey defeats Truman and cancer statistics. Journal of the National Cancer Institute *101*, 1157. 10.1093/jnci/djp203.

Friedenson, B. (2013). Mutations in components of antiviral or microbial defense as a basis for breast cancer. Funct Integr Genomics *13*, 411-424. 10.1007/s10142-013-0336-1.

Friedenson, B. (2015). Mutations in Breast Cancer Exome Sequences Predict Susceptibility to Infections and Converge on the Same Signaling Pathways Journal of Genomes and Exomes J Genomes and Exomes *4*, 1-28.

Germini, D., Sall, F.B., Shmakova, A., Wiels, J., Dokudovskaya, S., Drouet, E., and Vassetzky, Y. (2020). Oncogenic Properties of the EBV ZEBRA Protein. Cancers (Basel) *12*. 10.3390/cancers12061479.

Hafez, A.Y., and Luftig, M.A. (2017). Characterization of the EBV-Induced Persistent DNA Damage Response. Viruses *9*. 10.3390/v9120366.

Hau, P.M., Lung, H.L., Wu, M., Tsang, C.M., Wong, K.L., Mak, N.K., and Lo, K.W. (2020). Targeting Epstein-Barr Virus in Nasopharyngeal Carcinoma. Front Oncol *10*, 600. 10.3389/fonc.2020.00600.

Helman, E., Lawrence, M.S., Stewart, C., Sougnez, C., Getz, G., and Meyerson, M. (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. Genome research *24*, 1053-1063. 10.1101/gr.163659.113.

Hu, H., Luo, M.L., Desmedt, C., Nabavi, S., Yadegarynia, S., Hong, A., Konstantinopoulos, P.A., Gabrielson, E., Hines-Boykin, R., Pihan, G., et al. (2016). Epstein-Barr Virus Infection of Mammary Epithelial Cells Promotes Malignant Transformation. EBioMedicine *9*, 148-160. 10.1016/j.ebiom.2016.05.025.

Kim, K.D., Tanizawa, H., De Leo, A., Vladimirova, O., Kossenkov, A., Lu, F., Showe, L.C., Noma, K.I., and Lieberman, P.M. (2020). Epigenetic specifications of host chromosome docking sites for latent Epstein-Barr virus. Nat Commun *11*, 877. 10.1038/s41467-019-14152-8.

Kioussis, D., and Pachnis, V. (2009). Immune and nervous systems: more than just a superficial similarity? Immunity *31*, 705-710. 10.1016/j.immuni.2009.09.009.

Kumar, R., Nagpal, G., Kumar, V., Usmani, S.S., Agrawal, P., and Raghava, G.P.S. (2019). HumCFS: a database of fragile sites in human chromosomes. BMC genomics *19*, 985. 10.1186/s12864-018-5330-5.

Latour, S., and Winter, S. (2018). Inherited Immunodeficiencies With High Predisposition to Epstein-Barr Virus-Driven Lymphoproliferative Diseases. Frontiers in immunology *9*, 1103. 10.3389/fimmu.2018.01103.

Law, E.K., Levin-Klein, R., Jarvis, M.C., Kim, H., Argyris, P.P., Carpenter, M.A., Starrett, G.J., Temiz, N.A., Larson, L.K., Durfee, C., et al. (2020). APOBEC3A catalyzes mutation and drives carcinogenesis in vivo. The Journal of experimental medicine *217*. 10.1084/jem.20200261.

Lawson, J.S., and Glenn, W.K. (2021). Catching viral breast cancer. Infectious agents and cancer *16*, 37. 10.1186/s13027-021-00366-3.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Leibowitz, M.L., Zhang, C.Z., and Pellman, D. (2015). Chromothripsis: A New Mechanism for Rapid Karyotype Evolution. Annu Rev Genet *49*, 183-211. 10.1146/annurev-genet-120213-092228.
Lopez, C., Kleinheinz, K., Aukema, S.M., Rohde, M., Bernhart, S.H., Hubschmann, D., Wagener, R., Toprak, U.H., Raimondi, F., Kreuz, M., et al. (2019). Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. Nat Commun *10*, 1459. 10.1038/s41467-019-08578-3.
Lorenzetti, M.A., De Matteo, E., Gass, H., Martinez Vazquez, P., Lara, J., Gonzalez, P., Preciado, M.V., and Chabay, P.A. (2010). Characterization of Epstein Barr virus latency pattern in Argentine breast carcinoma. PloS one *5*, e13603. 10.1371/journal.pone.0013603.
Lu, F., Wikramasinghe, P., Norseen, J., Tsai, K., Wang, P., Showe, L., Davuluri, R.V., and Lieberman, P.M. (2010). Genome-wide analysis of host-chromosome binding sites for Epstein-Barr Virus Nuclear Antigen 1 (EBNA1). Virology journal *7*, 262. 10.1186/1743-422X-7-262.
Lynn, D.J., Winsor, G.L., Chan, C., Richard, N., Laird, M.R., Barsky, A., Gardy, J.L., Roche, F.M., Chan, T.H., Shah, N., et al. (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. Molecular systems biology *4*, 218. 10.1038/msb.2008.55.
Maccaroni, K., Balzano, E., Mirimao, F., Giunta, S., and Pelliccia, F. (2020). Impaired Replication Timing Promotes Tissue-Specific Expression of Common Fragile Sites. Genes (Basel) *11*. 10.3390/genes11030326.
Marrao, G., Habib, M., Paiva, A., Bicout, D., Fallecker, C., Franco, S., Fafi-Kremer, S., Simoes da Silva, T., Morand, P., Freire de Oliveira, C., and Drouet, E. (2014). Epstein-Barr virus infection and clinical outcome in breast cancer patients correlate with immune cell TNF-alpha/IFN-gamma response. BMC cancer *14*, 665. 10.1186/1471-2407-14-665.
McClintock, B. (1941). The stability of broken ends of chromosomes in Zea Mays. Genetics *26*, 234-282.
Moquin, S.A., Thomas, S., Whalen, S., Warburton, A., Fernandez, S.G., McBride, A.A., Pollard, K.S., and Miranda, J.L. (2018). The Epstein-Barr Virus Episome Maneuvers between Nuclear Chromatin Compartments during Reactivation. Journal of virology *92*. 10.1128/JVI.01413-17.
Mount, D.W. (2007). Using the Basic Local Alignment Search Tool (BLAST). CSH Protoc *2007*, pdb top17. 10.1101/pdb.top17.
Nakshatri, H., Bhat-Nakshatri, P., Martin, D.A., Goulet, R.J., Jr., and Sledge, G.W., Jr. (1997). Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth. Molecular and cellular biology *17*, 3629-3639. 10.1128/MCB.17.7.3629.
Needleman, S.B.a.W., C.D. (1970). A general method applicable to search for similarities in the amino acid sequence of two proteins. Journal of molecular biology *48*, 453-453.
Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *534*, 47-54. 10.1038/nature17676.
Nik-Zainal, S., Davies, H., Staaf, J., Ramakrishna, M., Glodzik, D., Zou, X., Martincorena, I., Alexandrov, L.B., Martin, S., Wedge, D.C., et al. (2019). Author Correction: Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature *566*, E1. 10.1038/s41586-019-0883-2.
Nik-Zainal, S., and Morganella, S. (2017). Mutational Signatures in Breast Cancer: The Problem at the DNA Level. Clinical cancer research : an official journal of the American Association for Cancer Research *23*, 2617-2629. 10.1158/1078-0432.CCR-16-2810.
Nones, K., Johnson, J., Newell, F., Patch, A.M., Thorne, H., Kazakoff, S.H., de Luca, X.M., Parsons, M.T., Ferguson, K., Reid, L.E., et al. (2019). Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. Annals of oncology : official journal of the European Society for Medical Oncology / ESMO *30*, 1071-1079. 10.1093/annonc/mdz132.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Ortutay, C., Siermala, M., and Vihinen, M. (2007a). ImmTree: database of evolutionary relationships of genes and proteins in the human immune system. Immunome Res *3*, 4. 10.1186/1745-7580-3-4.
Ortutay, C., Siermala, M., and Vihinen, M. (2007b). Molecular characterization of the immune system: emergence of proteins, processes, and domains. Immunogenetics *59*, 333-348. 10.1007/s00251-007-0191-0.
Peddu, V., Dubuc, I., Gravel, A., Xie, H., Huang, M.L., Tenenbaum, D., Jerome, K.R., Tardif, J.C., Dube, M.P., Flamand, L., and Greninger, A.L. (2019). Inherited Chromosomally Integrated Human Herpesvirus 6 Demonstrates Tissue-Specific RNA Expression In Vivo That Correlates with an Increased Antibody Immune Response. J Virol *94*. 10.1128/JVI.01418-19.
Peng, J., Wang, T., Zhu, H., Guo, J., Li, K., Yao, Q., Lv, Y., Zhang, J., He, C., Chen, J., et al. (2014). Multiplex PCR/mass spectrometry screening of biological carcinogenic agents in human mammary tumors. Journal of clinical virology : the official publication of the Pan American Society for Clinical Virology *61*, 255-259. 10.1016/j.jcv.2014.07.010.
Prusty, B.K., zur Hausen, H., Schmidt, R., Kimmel, R., and de Villiers, E.M. (2008). Transcription of HERV-E and HERV-E-related sequences in malignant and non-malignant human haematopoietic cells. Virology *382*, 37-45. 10.1016/j.virol.2008.09.006.
Scott, R.S. (2017). Epstein-Barr virus: a master epigenetic manipulator. Curr Opin Virol *26*, 74-80. 10.1016/j.coviro.2017.07.017.
Shapiro, R.L. (1965). An analysis of variance test for normality (complete samples). Biometrika *52*, 20.
Sinclair, A.J., Moalwi, M.H., and Amoaten, T. (2021). Is EBV Associated with Breast Cancer in Specific Geographic Locations? Cancers (Basel) *13*. 10.3390/cancers13040819.
Spehalski, E., Kovalchuk, A.L., Collins, J.T., Liang, G., Dubois, W., Morse, H.C., 3rd, Ferguson, D.O., Casellas, R., and Dunnick, W.A. (2012). Oncogenic Myc translocations are independent of chromosomal location and orientation of the immunoglobulin heavy chain locus. Proceedings of the National Academy of Sciences of the United States of America *109*, 13728-13732. 10.1073/pnas.1202882109.
Staaf, J., Glodzik, D., Bosch, A., Vallon-Christersson, J., Reutersward, C., Hakkinen, J., Degasperi, A., Amarante, T.D., Saal, L.H., Hegardt, C., et al. (2019). Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nature medicine *25*, 1526-1533. 10.1038/s41591-019-0582-4.
Tang, M.H., Varadan, V., Kamalakaran, S., Zhang, M.Q., Dimitrova, N., and Hicks, J. (2012). Major chromosomal breakpoint intervals in breast cancer co-localize with differentially methylated regions. Front Oncol *2*, 197. 10.3389/fonc.2012.00197.
Tsurumi, T., Kobayashi, A., Tamai, K., Yamada, H., Daikoku, T., Yamashita, Y., and Nishiyama, Y. (1996). Epstein-Barr virus single-stranded DNA-binding protein: purification, characterization, and action on DNA synthesis by the viral DNA polymerase. Virology *222*, 352-364. 10.1006/viro.1996.0432.
Tweedy, J., Spyrou, M.A., Pearson, M., Lassner, D., Kuhl, U., and Gompels, U.A. (2016). Complete Genome Sequence of Germline Chromosomally Integrated Human Herpesvirus 6A and Analyses Integration Sites Define a New Human Endogenous Virus with Potential to Reactivate as an Emerging Infection. Viruses *8*. 10.3390/v8010019.
Umbreit, N.T., Zhang, C.Z., Lynch, L.D., Blaine, L.J., Cheng, A.M., Tourdot, R., Sun, L., Almubarak, H.F., Judge, K., Mitchell, T.J., et al. (2020). Mechanisms generating cancer genome complexity from a single cell division error. Science *368*. 10.1126/science.aba0712.
Wolfe, D., Dudek, S., Ritchie, M.D., and Pendergrass, S.A. (2013). Visualizing genomic information across chromosomes with PhenoGram. BioData Min *6*, 18. 10.1186/1756-0381-6-18.

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Wu, C.C., Liu, M.T., Chang, Y.T., Fang, C.Y., Chou, S.P., Liao, H.W., Kuo, K.L., Hsu, S.L., Chen, Y.R., Wang, P.W., et al. (2010). Epstein-Barr virus DNase (BGLF5) induces genomic instability in human epithelial cells. Nucleic acids research *38*, 1932-1949. 10.1093/nar/gkp1169.
Xiao, K., Yu, Z., Li, X., Li, X., Tang, K., Tu, C., Qi, P., Liao, Q., Chen, P., Zeng, Z., et al. (2016). Genome-wide Analysis of Epstein-Barr Virus (EBV) Integration and Strain in C666-1 and Raji Cells. Journal of Cancer *7*, 214-224. 10.7150/jca.13150.
Xu, M., Yao, Y., Chen, H., Zhang, S., Cao, S.M., Zhang, Z., Luo, B., Liu, Z., Li, Z., Xiang, T., et al. (2019a). Genome sequencing analysis identifies Epstein-Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. Nature genetics *51*, 1131-1136. 10.1038/s41588-019-0436-5.
Xu, M., Zhang, W.L., Zhu, Q., Zhang, S., Yao, Y.Y., Xiang, T., Feng, Q.S., Zhang, Z., Peng, R.J., Jia, W.H., et al. (2019b). Genome-wide profiling of Epstein-Barr virus integration by targeted sequencing in Epstein-Barr virus associated malignancies. Theranostics *9*, 1115-1124. 10.7150/thno.29622.
Zapatka, M., Borozan, I., Brewer, D.S., Iskar, M., Grundhoff, A., Alawi, M., Desai, N., Sultmann, H., Moch, H., Pathogens, P., et al. (2020). The landscape of viral associations in human cancers. Nature genetics *52*, 320-330. 10.1038/s41588-019-0558-9.
Zhang, X., Tan, Y., Ling, Y., Lu, G., Liu, F., Yi, Z., Jia, X., Wu, M., Shi, B., Xu, S., et al. (2020). Viral and host factors related to the clinical outcome of COVID-19. Nature *583*, 437-440. 10.1038/s41586-020-2355-0.
Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol *7*, 203-214. 10.1089/10665270050081478.

Supplementary Table S1 Absence of inverted repeats at breakpoints in BRCA2 associated breast cancers. Chromosome coordinates for breaks vs nearby unbroken sequences were assayed for repeats within 100 base pairs in either direction using “RepeatAround”.

Status	Breakpoint or non-Breakpoint	Direct Repeat	Inverted Repeat	Mirror Repeat	Complementary Repeat
No Breaks	1:105993316	2 (1 8bps, 1 9bps)	0	0	0
Breaks	1:102731470	1 (8 bps)	0	0	0
Breaks	1:104326329	1	0	0	0
Breaks	1:145685562	1	0	0	0
No Breaks	1:143999000	2 (8,10bps)	0	2 (8 bps)	0
No Breaks	2:23000000	4 (8,8,10,10 bps)	0	0	0
Break	2: 25,505,554	1 (10 bps)	0	0	0
Breaks	2:100,035,528	1(10 bps)	0	0	0
Breaks	4:101,009,819	2(8,8 bps)	0	0	0
No breaks	4:102,639,952	2(9, 20 bps)	0	0	0
Breaks	8:80,687,247	2 (8,14 bps)	0	0	0
No Breaks	8:83,500,000	2 (8,8 bps)	0	0	0

bioRxiv preprint doi: https://doi.org/10.1101/2021.11.08.467751; this version posted November 28, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Breaks	11:94,386,526	2(9, 13 bps)	0	0	0
No Breaks	11:91,962,848	6(8,8,8,9,12)	0	0	0
Breaks	12:88,191,644	1(14 bps)	0	0	0
No Breaks	12:90,932,271	1(12 bps)	0	0	0