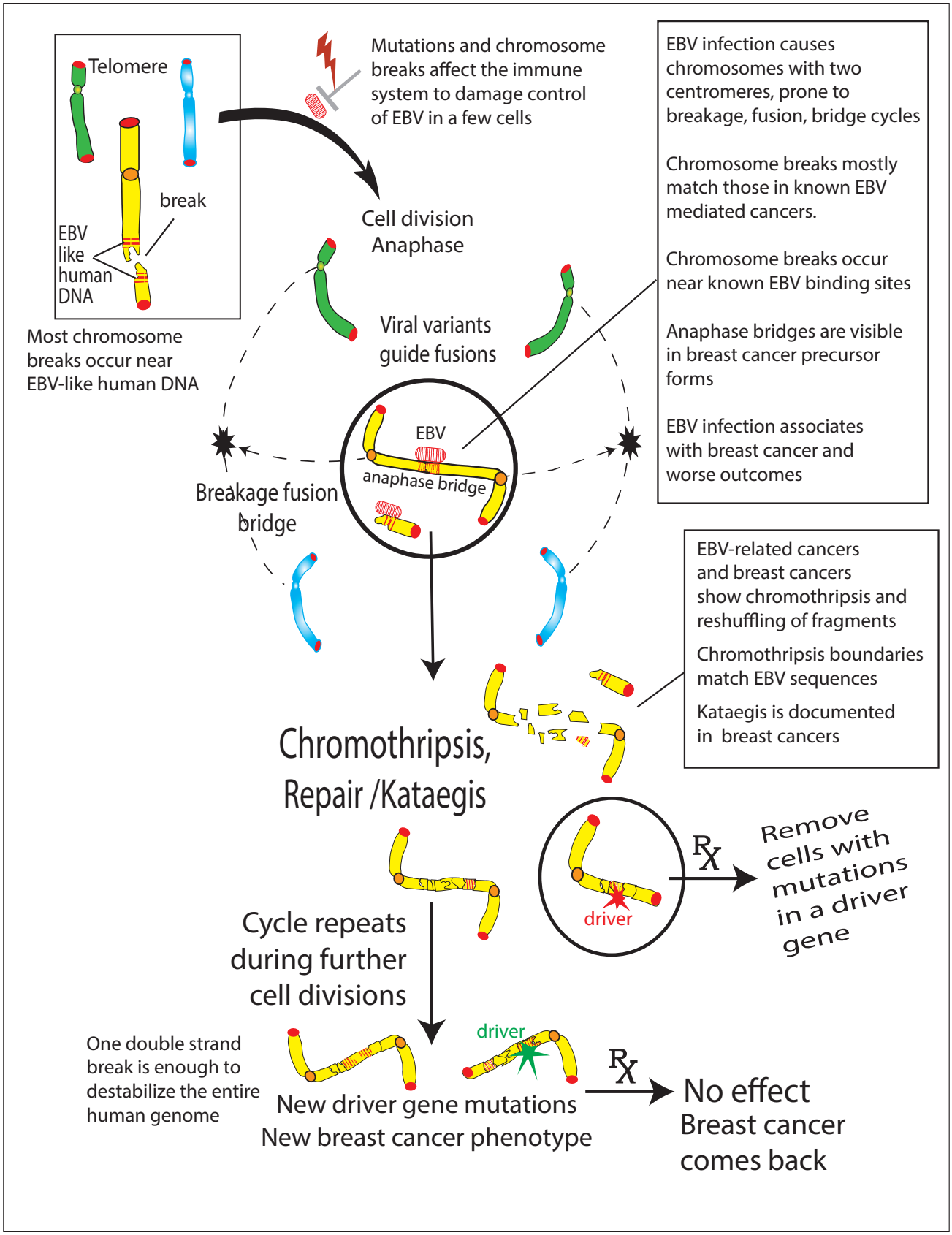
Chromosome breaks in breast cancers occur near herpes tumor virus sequences and explain why the cancer comes back

Bernard Friedenson

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

Correspondence: bernief@UIC.edu

Graphical Abstract



Abstract

One explanation for why breast cancer has a relentless tendency to come back after treatment is that breast cancer phenomes are fluid, so it is difficult for therapy to target new phenotypes that arise. In breast cancer, viral interactions cause chromosome breakage and fusion, producing anaphase bridges containing two centromeres. Cell division then leads to more chromosome breaks, more mutations, and new cancer phenotypes. Analyses of public data from about 2100 breast cancers implicate Epstein-Barr virus (EBV or Human Herpes Virus 4) in the cycles that break chromosomes and produce anaphase bridges. EBV generates breakage cycles with highly fragmented chromosomes (chromothripsis) as an inherent feature. These episodes continually produce new cancer driver mutations. DNA at chromothripsis boundaries in breast cancers consistently matches EBV tumor variants. Breakpoints are often near known EBV binding sites, and breast cancer shares an EBV methylation signature with other known EBV cancers. Breast cancer breakpoints on all chromosomes cluster around the same positions as in cancers definitively associated with EBV infection (nasopharyngeal cancer and Burkitt's lymphoma). The status of EBV in the microbiome and the presence of anaphase bridges are phenotypic markers of breast cancer that are not explicitly targeted by chemotherapy. The breast cancer phenome continually supplies new cancer driver genes that foster breast cancer recurrence.

Summary

Breaks in breast cancer chromosomes and in fragments of anaphase bridges occur mainly near regions that resemble Epstein-Barr virus (EBV) tumor variants

Extensive breaks are an inherent outcome of anaphase bridges and the reassembly of highly fragmented chromosomes continually generates new cancer driver mutations and new cancer phenotypes.

Treatment does not explicitly target the underlying breakage and reassembly cycles so breast cancer often comes back.

The breast cancer phenome is not fixed and can continuously generate new phenotypes

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

Introduction

Breast cancer often comes back after therapy with unrelenting recurrence rates over at least 20 years. Depending on the number of lymph nodes involved, recurrence rates range from 12% (0 nodes) to 41% (4-9 nodes) (Pan *et al.*, 2017). One explanation for this recurrence is that treatment has overlooked an underlying cause of the disease. By the time breast cancer is diagnosed, its causes are difficult to isolate from multiple risk factors, heredity, thousands of mutations, and a lifetime of exposure to mutagens and infections.

Human cancer viruses represent risk factors that usually just initiate or promote cancer and do not cause cancer by themselves (Sarid and Gao 2011). For example, Epstein-Barr virus /Human herpesvirus 4 (EBV/HHV4) exists as a latent infection in almost all humans (>90%). Most people control EBV infection (Fina *et al.* 2001; Lawson and Glenn 2021), Life-long infections of EBV in most people obscures potential associations with breast cancer in a few people, but some associations are undeniable. 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. At age 18-19, 89% of the population is seropositive (Balfour *et al.* 2013).

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). In addition, EBV activation from its latent state causes massive changes in host chromatin methylation and structure (Buschle *et al.* 2021; 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). EBV can also trans-activate endogenous retroviruses (Bruce *et al.* 2021; Mameli *et al.* 2012; Meier *et al.* 2021),

In B-lymphocytes, EBV causes chromosome structural changes that foster breakage fusion-bridge-cycles and characteristic chromosome rearrangements. These structural changes include forming chromosomes with two centromeres (dicentric chromosomes), chromosomal deletions, fragmentation, unbalanced translocations, and telomere dysfunction (Kamranvar *et al.* 2007; Lacoste *et al.* 2010). Chromosomes with two centromeres are significantly increased in EBV positive Burkitt lymphoma cells (Kamranvar *et al.* 2007).

The present study determines whether nearly universal human EBV infection is associated with the same kinds of breast cancer chromosome breakages seen in EBV mediated tumors. The work grew from the observation that breast cancers have the same kinds of chromosome damage seen in cancers with established EBV relationships. These model EBV-related cancers have increases in chromosome breaks, rearrangements, fusions, deletions, and insertions (Cuceu *et al.* 2018). Even a single break can cause breakage, fusion, and bridging. These three events then repeat during every cell division to destabilize the entire human genome, producing many further complex rearrangements. Chromothripsis occurs in all cancers much more frequently than previously thought (Cortes-Ciriano *et al.* 2020) inherent to breakage, fusion and bridge cycles (Umbreit *et al.* 2020). The origins of chromothripsis are unknown and X-radiation is certainly a possible contributor (Stephens *et al.* 2011). Radiation therapy of breast cancer patients generates chromosomes with two centromeres in peripheral blood (Lee *et al.* 2021)

To investigate the roots of breast cancer breaks, the present study starts with nasopharyngeal cancer (NPC) as a model for an

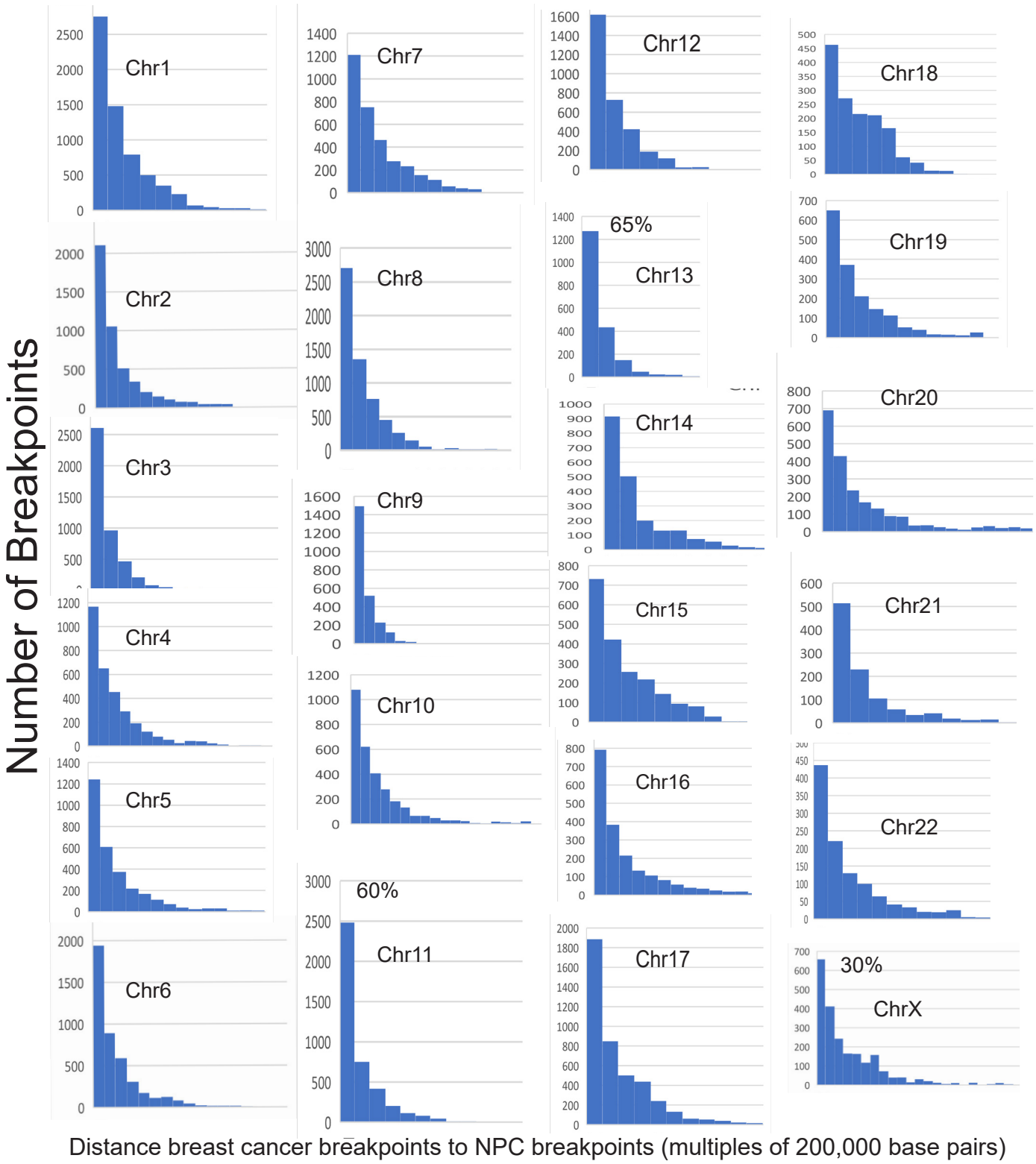
EBV-associated epithelial cell cancer. NPC has an explicit EBV connection with 100% of malignant cells being EBV-positive (Germini *et al.* 2020; Hau *et al.* 2020; Xu *et al.* 2019). The viral genome in NPC had nearly 8500 EBV forms in patients with over 2100 variants in a single host, each differing only slightly from a reference viral genome (Xu *et al.* 2019). Variants of BALF2, a viral DNA binding protein active during the lytic phase, are significantly associated with NPC (Tsurumi *et al.* 1996; Xu *et al.* 2019). NPC host mutations affect innate immunity and constitutively activate an inflammatory response (Bruce *et al.* 2021). These changes release controls on EBV infection. The level of EBV DNA in plasma after treatment predicts distant metastasis (Li *et al.* 2017). Similarly in breast cancer, the replicative form of EBV correlates with poorer outcomes. Latent infection confers a survival advantage and accompanies anti-tumoral immune responses (Marrao et al. 2014).

In breast cancer as in NPC, mutations or downregulation in DNA repair genes linked to *BRCA1-BRCA2* compromise immunity (Friedenson, 2013). Immune deficits in sporadic breast cancers are well-known. Powerful coordinated whole genome sequencing studies have delved deeply into the characteristics and origins of cancer genomes (Aleksandrov *et al.* 2020; Batra *et al.* 2021; Bruce *et al.* 2015; Cortes-Ciriano *et al.* 2020; Li *et al.* 2020; Nik-Zainal *et al.* 2016; Nones *et al.* 2019; Ou, 2020; Stephens *et al.* 2012; Voronina *et al.* 2020). These landmark studies provide robust starting points for the present work and they are extended here to compare breakpoints in breast cancers to breakpoints in cancers with proven relationships to herpes viruses.

Results

Based on genomes from 139 hereditary breast cancers and 70 nasopharyngeal cancers (NPC), every human chromosome in female breast cancers has breakpoints that cluster near those in NPC, a known EBV-mediated cancer (Fig. 1). Every chromosome is divided into 200,000 base pair segments and the number of breakpoints in each segment is compared for breast cancers vs NPC.. The tall bars 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. Fig. 2 (top) shows Chromosome 1 as a more detailed example of breakpoints in breast cancer relative to NPC breakpoints. With few exceptions, breakpoints in the two unrelated data sets agree well. The data is not normalized or filtered in any way.

If chromosome breakpoint positions follow probability theory, they 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 find no Gaussian distribution for thousands of breakpoints in any hereditary breast cancer, sporadic breast cancer, or nasopharyngeal cancer on any chromosome. Instead, many breakpoint positions are approximately the same among different cancers. To test the possibility that the sharing of breakpoints is a statistical artifact, 3000 breakpoints in NPC and breast



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 200000 base pairs of a NPC breakpoint vary from 65% on Chr13 to 30% on ChrX. Each bar on the graph is separated by 200000 base pairs on the X axis

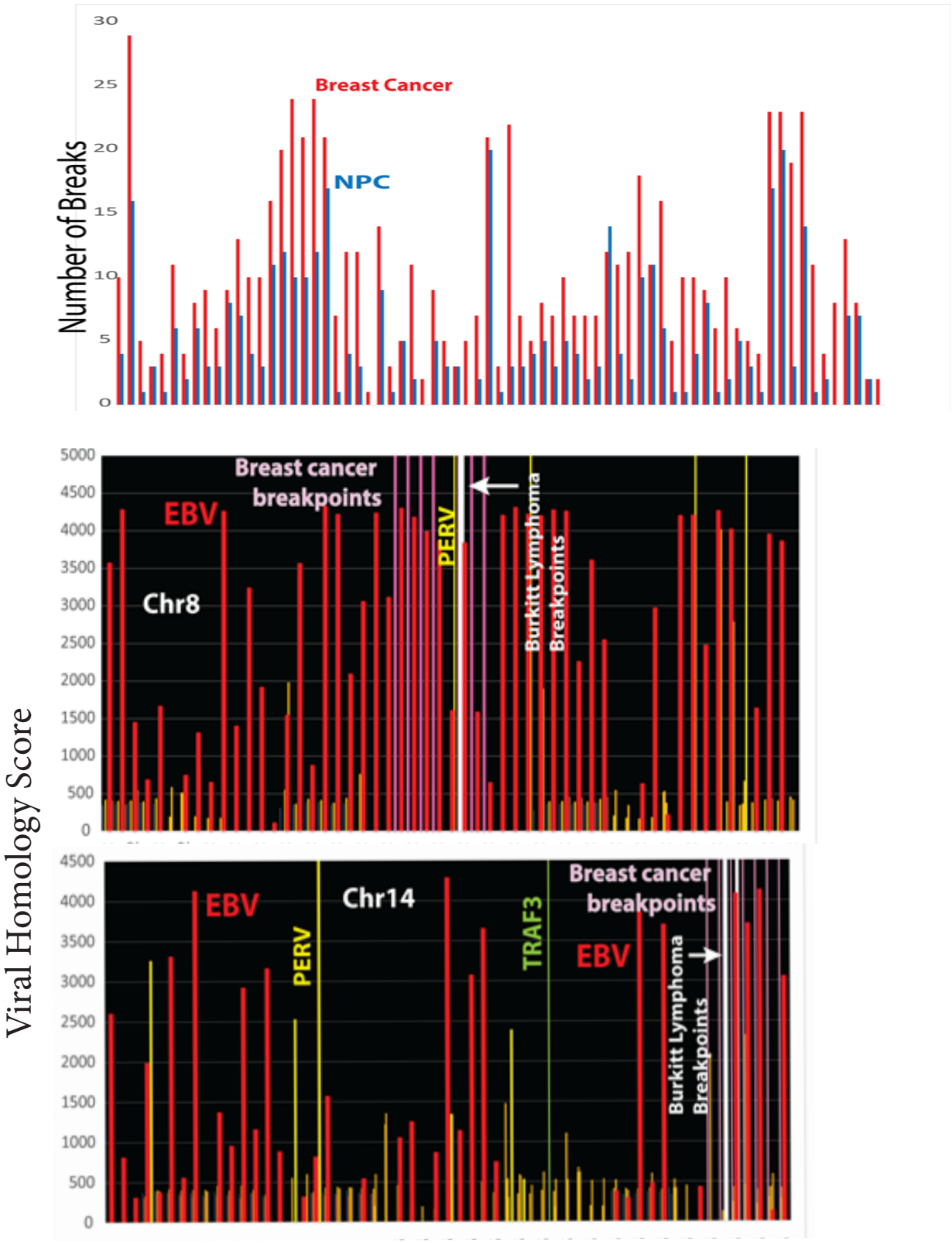


Fig. 2. Breakpoints at the MYC gene in Burkitt's lymphomas (white line), a cancer strongly associated with EBV are surrounded by DNA similar to EBV variants HKHD40 and HKNPC60 (indicated in red). Two areas of homology to porcine endogenous retrovirus (yellow lines) may also be relevant but are more distant

cancer were each represented by a set of random numbers between 1 and 150,000,000 (a typical chromosome length). The closest random "NPC' breakpoint 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 base pairs.

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

Breakpoints in Burkitt's lymphoma are near EBV-like sequences. Endemic Burkitt's lymphoma is another cancer strongly linked to EBV infection. Burkitt's lymphomas have gene rearrangements in the Myc gene on chromosome 8, with minor variations in different patients (Busch *et al.* 2007). Fig. 2 (middle panel) focuses the Chromosome 8 homology data on a 10-megabyte region around the Myc gene. Multiple DNA segments that strongly resemble EBV surround Myc breakpoints characteristic of Burkitt's lymphoma. In fact, 59 segments resembling viruses were within 200,000 base pairs of MYC breakpoints. Of these, 54 were from EBV-like sequences. A breakpoint cluster in Burkitt's lymphoma often accompanies MYC breakpoints. This reciprocal break 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, and the nearby TRAF3 gene controls the recombination. The lower panel in Fig. 2 shows that reciprocal breakpoints on chr14 within the antibody H-chain gene cluster (Lopez *et al.* 2019) are also near a region with 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 these genes often overlap those affected by breakpoints in NPC (Fig 3). 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 control.

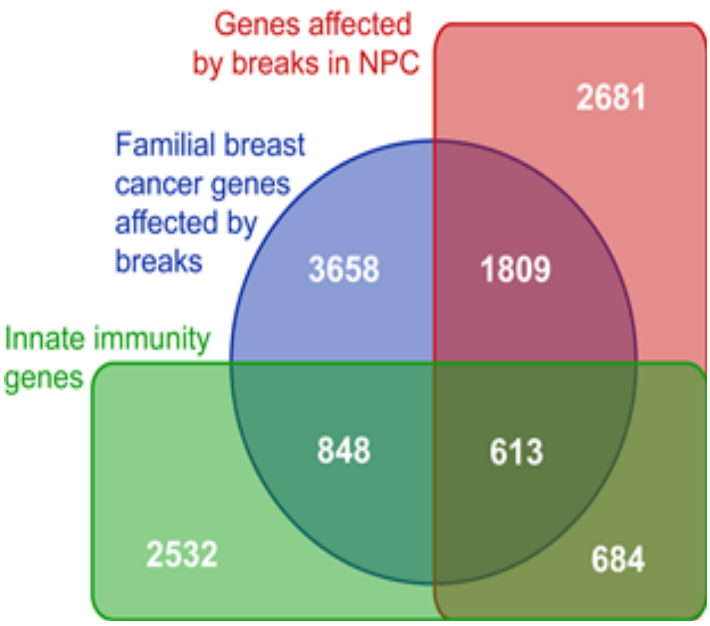


Fig. 3. Extensive overlap between genes affected by breakpoints among hereditary breast cancers, nasopharyngeal cancers and innate immunity.

Deregulation of innate immune responses increases mutagenesis and facilitates multiple human cancers (Law *et al.* 2020). A compromised or suppressed innate immune system allows reactivation and persistent infection by 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 NPC gene drivers: 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 inappro-

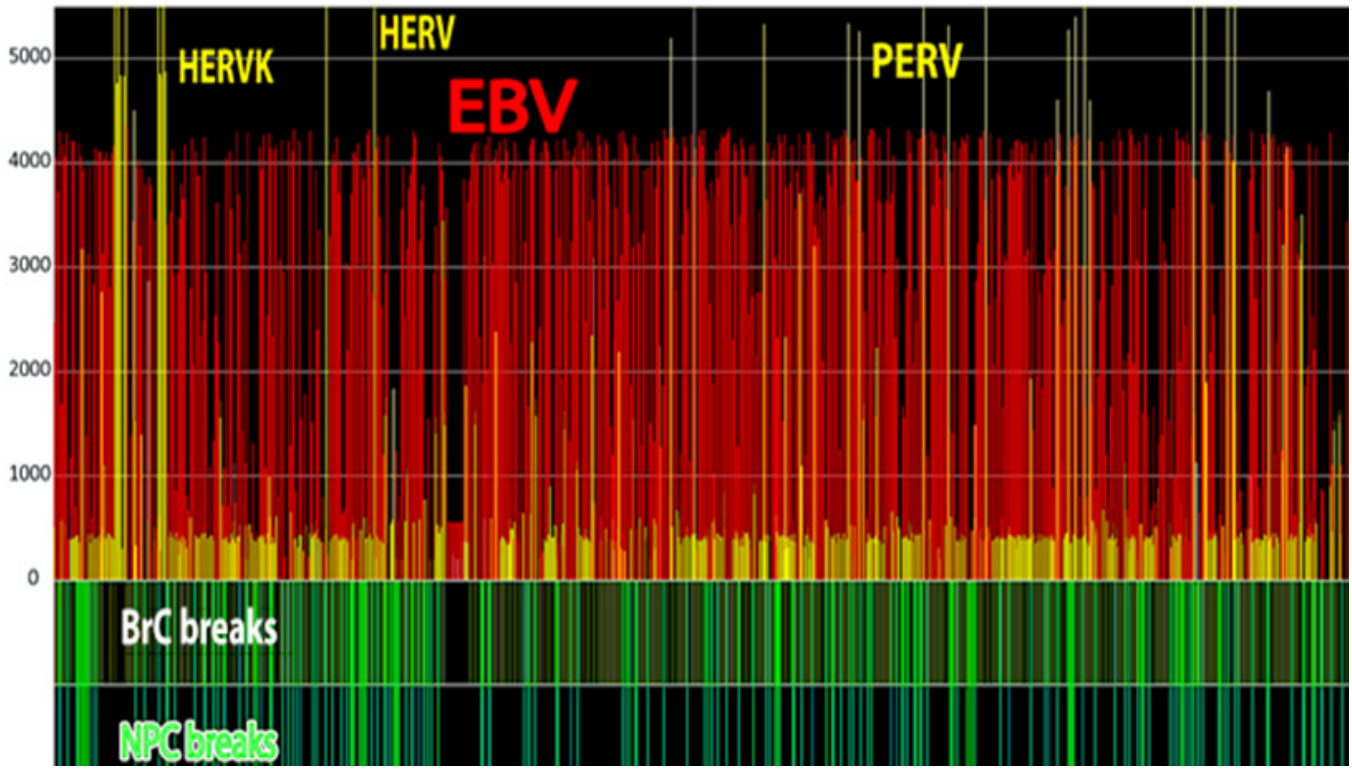


Fig. 4A. Maximum homology scores to all viruses vs. position on the entire length of chromosome 8. The strongest resemblance occurs to variants of only a few viruses indicated. EBV variants clearly predominate. Break positions in hereditary breast cancers and NPC are shown.

priate inflammatory response (Nakshatri *et al.* 1997). Other herpes-mediated cancers, such as Kaposi's sarcoma also occur in the context of compromised immunity.

Breakpoints occur near human sequences that resemble herpes viruses.

Virus-human homology comparisons were conducted around thousands of human BRCA1, BRCA2 and familial breast cancer-associated breakpoints using public data (Nik-Zainal *et al.* 2016; Nones *et al.* 2019). Long stretches of EBV variant DNA from two human gamma-herpesvirus 4 variants, HKN-PC60 and 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. Breakpoints in hereditary breast cancers cluster around human sequences that resemble herpes virus variants. Fig. 4A shows

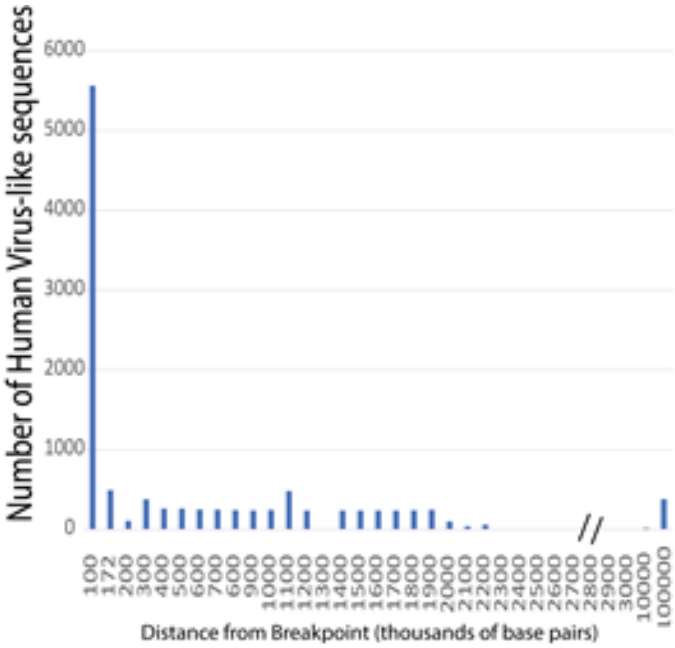


Fig. 4B. Most breakpoints on chr8 are within 100kb of a virus-like sequence.

all viral homologies on the 145,138,636 base pairs of 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 nonetheless strong.

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. 4b).

Breakpoints anywhere along the entire length of chromosome 11 were also compared to positions of homology to all viruses (Fig. S1). Chromosome 11 has 4141 inter- and intra-chromosomal breakpoints in 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 simulations of 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.

Fig. S1 also shows that EBV homologies 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 matches 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, 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, $p<0.0001$.

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. 5a), 63% of EBV docking sites (Kim *et al.* 2020) 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 EBV. Start positions of human genome similarity to the EBV variant tumor viruses HKNPC60 and HKHD40 positions aligned closely to almost all the estimated positions for 56 EBV docking sites (Fig. 5a). 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.

The lower panels in Figs. 5a and 5b support the idea that breakpoints occur in accessible regions and can have devastating effects. Breakpoints are all accessible as indicated by DNase hypersensitivity and transcription activity. Transcription factor footprints at these sites are an important predictor of complex traits and disease risks (Funk *et al.* 2020). 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

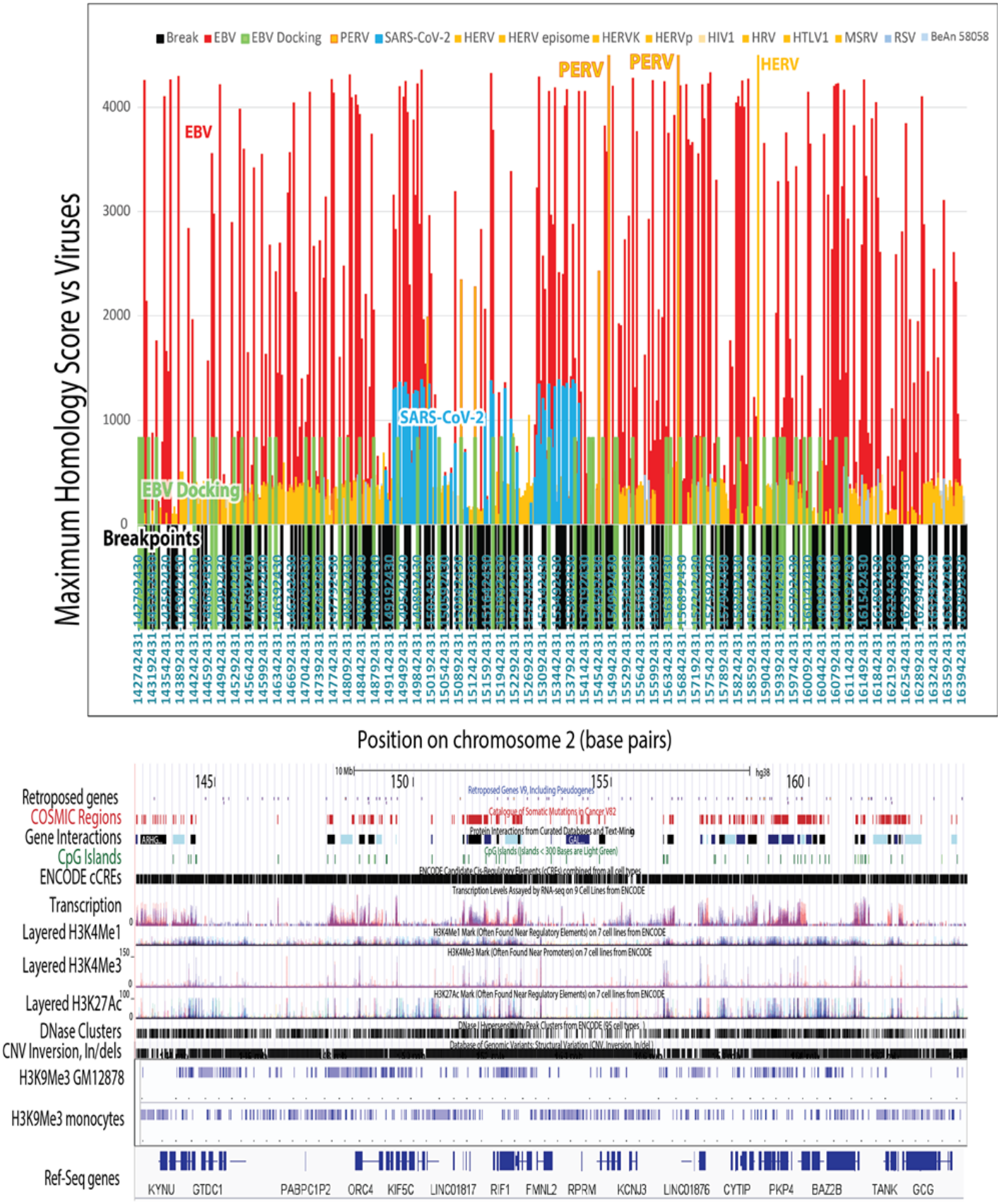


Fig. 5A. Breakpoints on a 21 million base pair segment in human chromosome 2 in 139 BRCA1, BRCA2, and presumptive BRCA-associated breast cancers. EBV anchor locations (green lines) and breast cancer breakpoints (black lines) are shown to allow comparisons. In the top panel, the axis at the left represents maximum homology to all viruses. The lower section shows characteristics of the human genome aligned with selected regions from the UCSC genome browser.

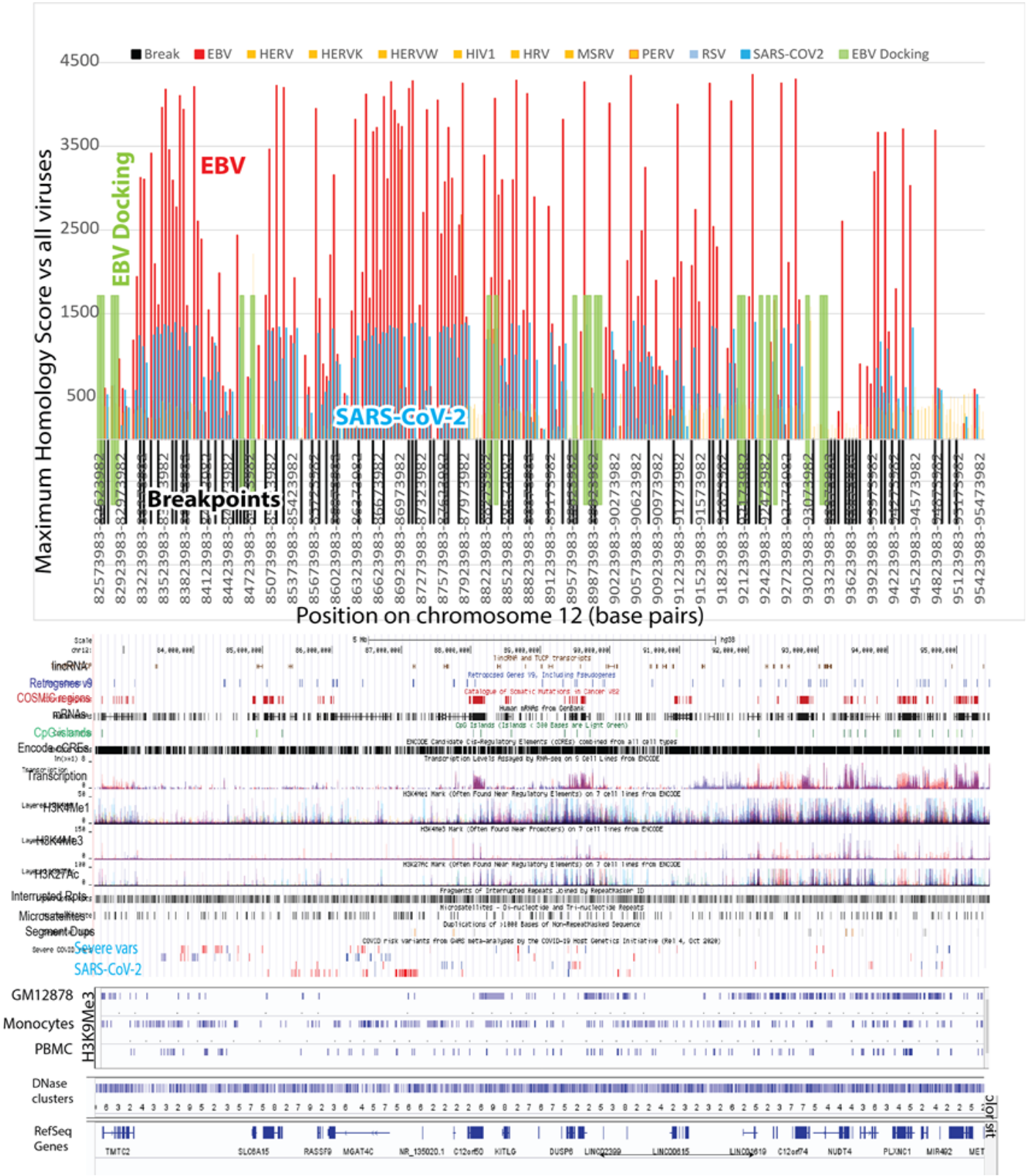


Fig. 5B. Breakpoints from 139 hereditary or likely hereditary breast cancers on a 13 million base pair segment of chromosome 12 are aligned with homology to all viruses and to selected parameters from the UCSC genome browser.

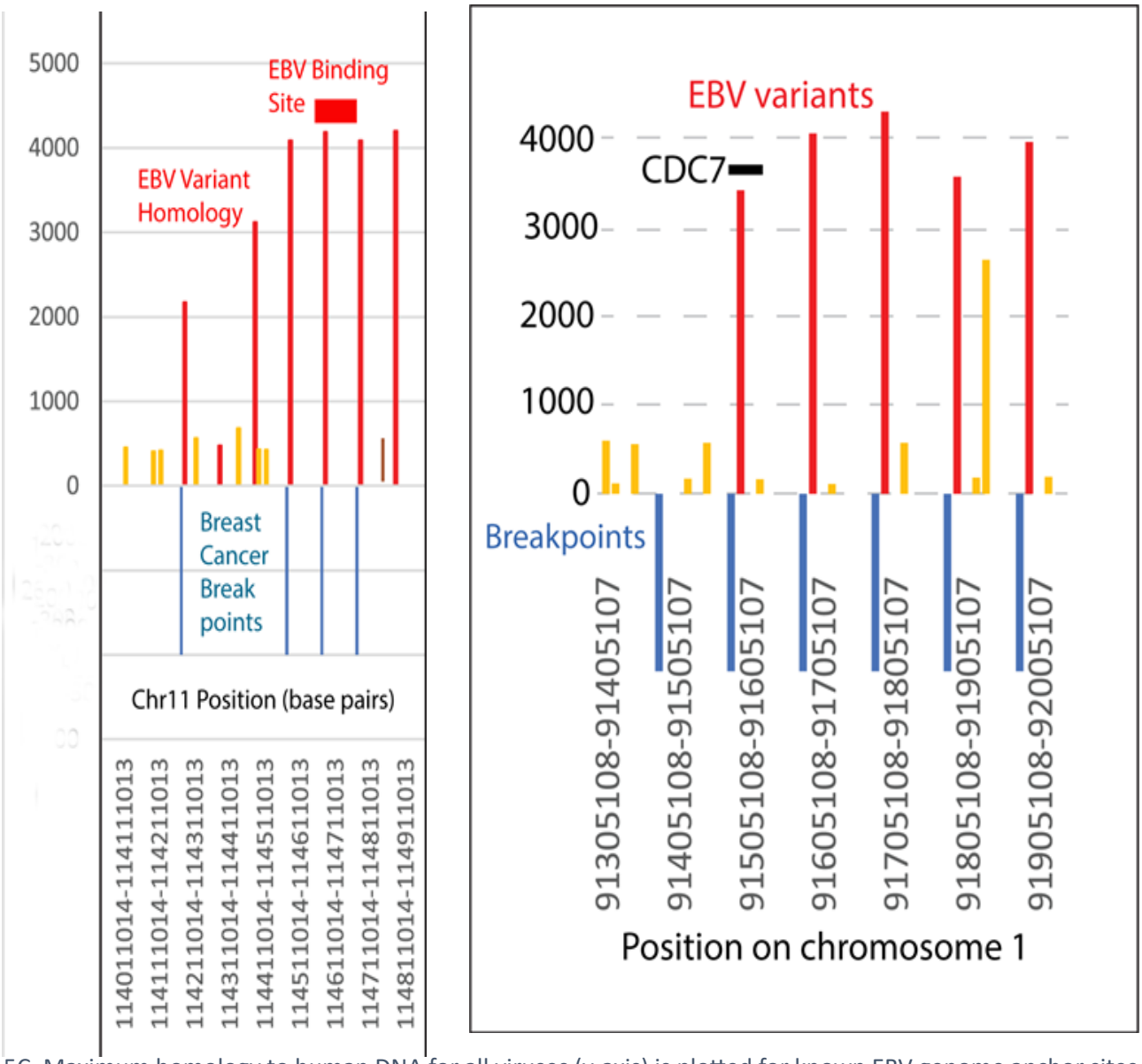


Fig. 5C. Maximum homology to human DNA for all viruses (y-axis) is plotted for known EBV genome anchor sites vs. breast cancer breakpoints and gene coordinates. At the left, a primary EBV binding site is close to several breast cancer breaks and human DNA that resembles EBV variants. At right, EBV variant 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.

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. 5a and 5b).

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

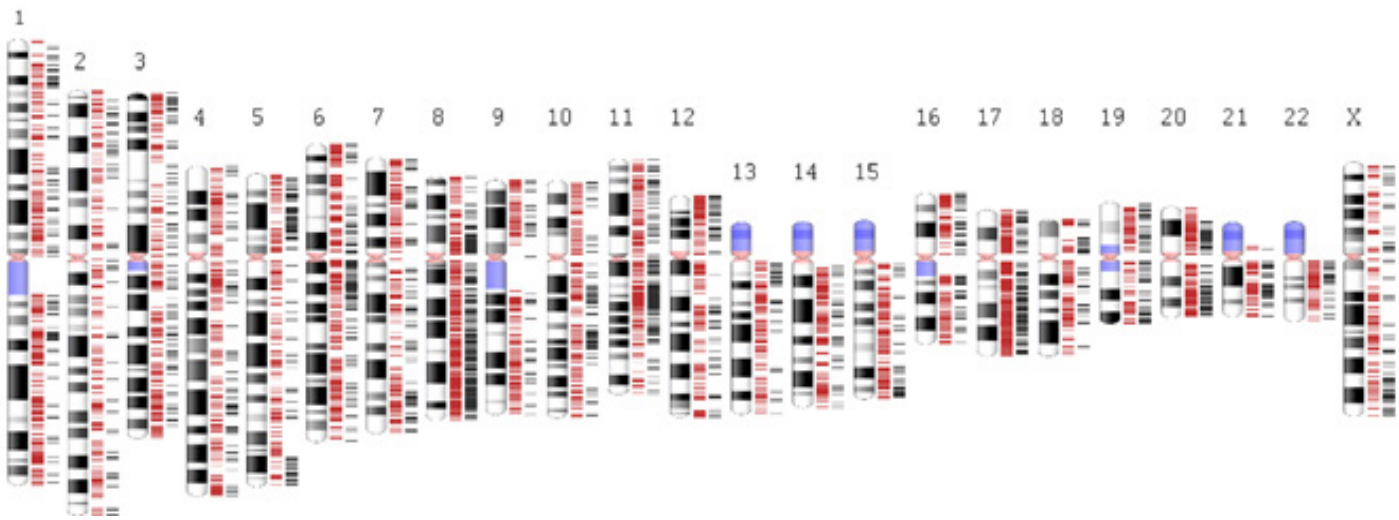


Fig. 6. Inter-chromosome translocation break positions in 74 mutation-associated or likely mutation associated female breast cancers (red) vs. 74 likely sporadic female breast cancers (black).

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. 5a, 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. 5b, bottom) include vesicle trafficking (*RASSF9*), endocytosis (*EEA1*), blood cell formation (*KITLG*), interferon response control (*SOCS2*), and nerve cell patterning (*NR2C1*).

Clusters of breakpoints are obvious in Fig 5A (black lines at the bottom of the graph in the upper figure panel). 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 (Supplementary Fig S2).

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

Fig. 5C 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 chr11 (Lu *et al.* 2010) closely matches a few breast cancer breakpoints. Another known EBV binding site on chr1 near the CDC7 gene corresponds to multiple breast cancer breaks.

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 mutations

(Nik-Zainal *et al.* 2016) were compared to breakpoints in 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. 6). Although there are significant differences in breakpoint distributions (Fig. 6), 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 marker predictive of BRCA1/BRCA2 deficiency (Staaf *et al.* 2019). However, triple-negative breast cancers comprised only 9% of breast cancers in the cohort.

Breakpoints in sporadic breast cancers on every chromosome are also 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 aberrant repair with kataegis. In addition, telomere dysfunction promotes end-to-end fusions. For 74 breast cancers in women over age 70 (Nik-Zainal *et al.* 2016) taken as sporadic cancers, 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 results from sporadic breast cancers (not shown) resemble the results for hereditary cancers (Fig.1). Generally, breakpoints are less common in sporadic cancers than in hereditary cancers, but sporadic cancer breakpoints are still related to NPC. On chromosome 11, sporadic breast cancer breakpoints are close to 78% of NPC breakpoints.

A potential EBV signature is present in sporadic breast cancers.

Although EBV-positive NPC and gastric cancers have distinctive patterns of DNA hypermethylation, some chromosome areas are hypermethylated in both cancers. 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. 7 (top) 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 (Fig 7, bottom).

Boundaries of chromothripsis regions align with EBV sequences.

Fig 8 shows regions that undergo chromothripsis with high confidence (Cortes-Ciriano *et al.* 2020) on the entire lengths of chromosome 6 (top) and chromosome 18 (middle).

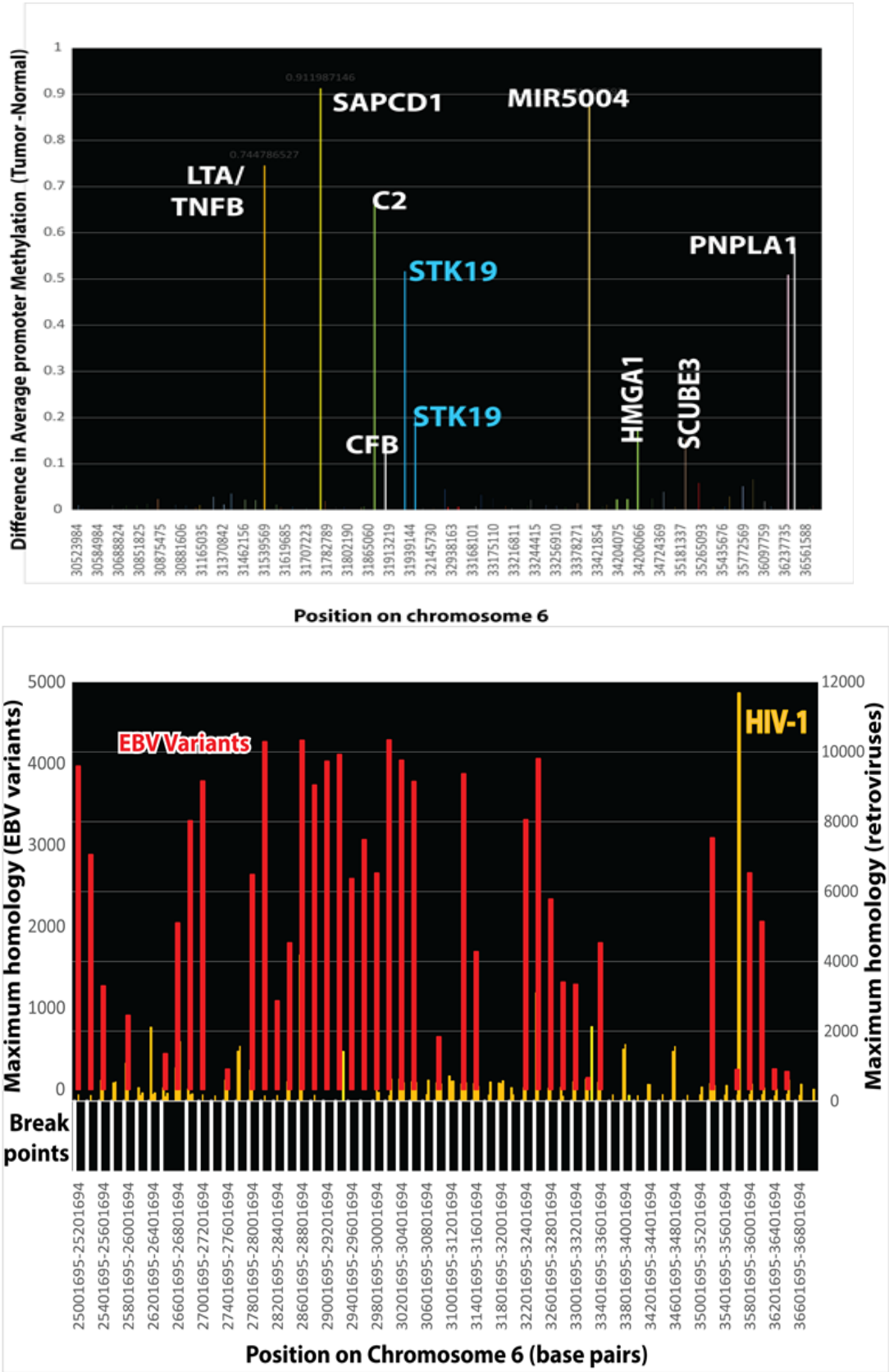


Fig. 7. Differences in methylation in 1538 breast cancers vs. normal controls in the potential EBV signature region on chromosome 6p21.3. The genes shown are primarily involved in complement pathways. Bottom. All but a few chromosome breakpoints in the 6p21.3 region are near areas of homology to EBV variants (red) or HIV1 or endogenous retroviral variants (orange or yellow)

All of the boundaries of these regions align with areas of homology to EBV. The bottom panel shows chromothripsis breakpoints on all chromosomes that are within about 200,000 bases of a region of viral homology. The boundaries of chromosome regions with high confidence chromothripsis breaks are primarily near long EBV-like sequences (1123 sequences or 57% of the total). Remaining chromothripsis breaks are near endogenous retroviruses (37%) or HIV-1 like sequences (9%). There are also strong matches to endogenous and other retroviruses. One breast cancer (donor 1347751) had several particularly strong matches to HERV-K (near chr11:101.6 megabases). The bottom panel shows viral homologies around chromothripsis boundaries on most chromosomes (white labels) with the numbers of base pairs separating regions of homology from chromothripsis boundaries.

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.* 2019). 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 has 99% and 98% identity in 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). 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 (Supplementary Fig. S3) and most other chromosomes are inconsistent with the fragile site database. In the PCAWG dataset, there are no fragile sites listed for chromosomes 8, 9, 11-15,17-19, 21, and 22 (Li *et al.* 2020). There are many more breakpoints than fragile sites on all the chromosomes. Breast cancer breaks do not consistently occur near common fragile sites. Some hereditary breast cancer breakpoints were tested for repeats likely to generate fragile site breaks because the repeats are prone to replication errors. This test did not find such sequences.

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 additional breaks, destabilizing the entire genome (Umbreit *et al.* 2020) and producing new cancer driver mutations. This cascade is especially likely in hereditary breast cancers with their deficits in homologous recombination repair.

Discussion

The present work shows human sequences related to tumor virus variants cluster around chromosome breaks in breast cancers. Virus like human sequences surround high-confidence chromothripsis breaks, an inherent feature of anaphase bridges. Multiple independent lines of evidence support this view and are summarized below (1-15).

1. Breakpoints in hereditary breast cancer, sporadic breast cancers, and NPC cluster around the same positions on all chromosomes in females. NPC is an epithelial cell cancer mediated by EBV variants.

2. EBV and endogenous retrovirus matching sequences occur near regions that undergo high-confidence chromothripsis in breast cancer. Chromothripsis is inherent in breakage-fusion-bridge cycles. Model EBV infected B-cells show chromosomes with two cen-

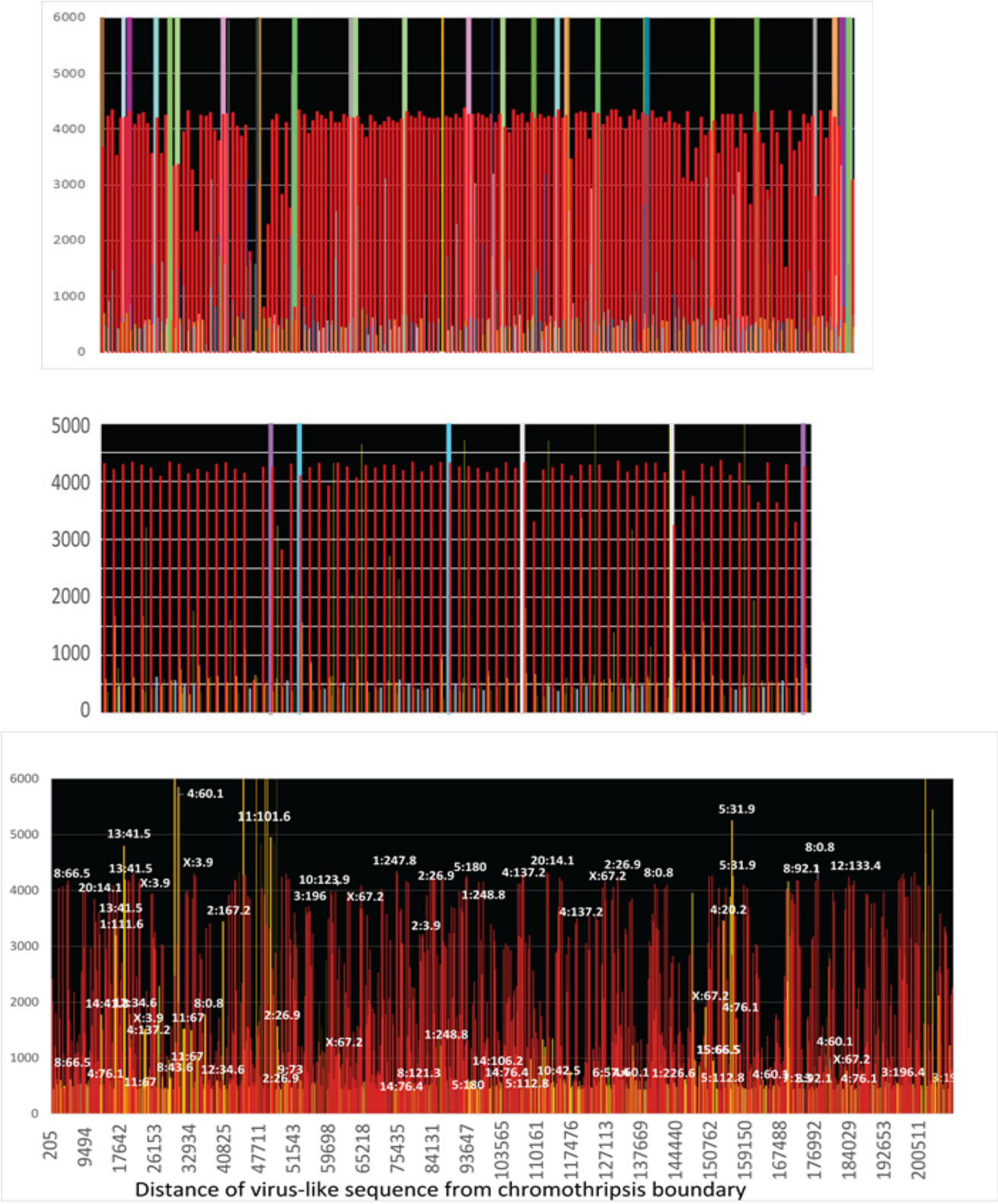


Fig. 8. Chromothripsis boundaries along the entire length of chr6 (top) and chr18 (middle) in different breast cancers align with regions of homology to EBV. Chromothripsis boundaries (Cortes-Ciriano et al., 2020) are indicated as thick colored or white lines, extending across the Y-axis. The thinner lines represent strong homology to human retroviruses and PERV. The bottom panel shows viral homologies around chromothripsis boundaries on most chromosomes (white text) with the numbers of base pairs separating regions of viral homology. Red bars indicate EBV-like sequences and orange indicate endogenous retroviruses.

tromeres and other evidence of breakage fusion bridge cycles.

3. Breast cancer breakpoints and EBV-like sequences on chromosome 14 cluster around the IgM class switch region characteristic of endemic Burkitt's lymphoma. Similarly, EBV-like DNA and a group of breast cancer breakpoints surround partner MYC region breakpoints in Burkitt's lymphoma.

4. A potential EBV methylation signature shared with known EBV-related cancers is far more abundant in 1538 breast cancers than in normal controls.

5. Breast cancer breakpoints are consistently near binding sites for EBV.

6. Every human chromosome in female breast cancers has breakpoints near EBV-like human DNA. EBV can also activate endogenous retroviral sequences, found near some breakpoints.

7. The association of breast cancer breaks and EBV tumor variants does not depend on the presence of BRCA1 or BRCA2 gene mutations. However, breast cancer breaks are more frequent in BRCA1 and BRCA2 mutation carriers than in non-carriers.

8. Sets of random breakpoints do not collect around similar values, like breakpoints in breast cancer and NPC.

9. 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 mutations also occur after illicit break repair.

10. 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.

11. Fragile site sequences are not sufficient to account for breast cancer breaks

12. Tumor variants HKHD40 and HKNPC60 are herpesviruses closely related to known tumor virus populations KHSV and EBV.

13. EBV infection is nearly universal and can cause genomic instability well before breast cancer occurs.

14. EBV can also activate endogenous retroviral sequences, which occur near some breakpoints.

15. 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.

Some of the most recent cancer drug therapy focuses on identifying and targeting cancer driver mutations. The drugs are initially effective, sometimes for long periods but then the drugs quit working. The graphical abstract summarizes a proposed model showing breakage-fusion-bridge cycles as occult, underlying processes that continue even if cells with the initial driver mutations are removed. It is known that cancer treatment generates new clones that do not exist in the original population (Aissa *et al.* 2021). The cancer phenome includes an inherent process that generates new cancer driver mutations and new cancer phenotypes (graphical abstract).

According to the model, mutations and chromosome breaks affecting immunity damage control of EBV infection. Chromosomes from populations of EBV variants infecting the cell bind to chromosome breaks in human DNA at matching sequences. The best conditions quickly repair breaks by rejoining the two fragments. However, repetitive virus-like human sequences fuse to inappropriate chromosome fragments.

A chromosome fragment is deficient in telomeres, so a second chromosome fragment can link to it, but the product can 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 characteristic of the process (Umbreit et al. 2020). This scenario occurs in EBV positive Burkitt lymphomas and in B-cells (Kamranvar *et al.* 2007; Lacoste *et al.* 2010) and produces visible anaphase bridges in precursor forms of breast cancer (Chin *et al.* 2004). Chromothripsis accompanying these bridges occurs in at least half of breast cancers (Cortes-Ciriano et al. 2020). Traditional chromothripsis involves alternate copy-number states for chromosome regions, but additional events involve multiple chromosomes and structural alterations. These alterations arise by non-homologous end joining, replication-difficulties and insertions of templated fragments. Chromothripsis can amplify oncogenes and inactivate DNA repair genes (Cortes-Ciriano 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. Effects related to APOBEC3 probably occur in EBV-induced carcinogenesis (Bobrovnitchaia et al. 2020; Law et al. 2020). The breakage fusion bridge process repeats during cell division, quickly increasing the number of abnormal chromosome products and driver mutations. When telomeres are removed or become so short that they cannot protect chromosome ends, APOBEC3B cytosine deamination (and perhaps TREX1 nucleolytic processing) cause chromothripsis and kataegis (Maciejowski et al. 2020; Umbreit et al. 2020). It is difficult to imagine how a spurious EBV-like chromosome would not contribute to these events.

Removing malignant cells by targeting a driver mutation stops cancer but this rarely involves 100% of cancer cells. As the remaining cells divide, the underlying process continues to produce more driver mutations and new phenotypes, so the cancer comes back. Once the process starts, it remains active without requiring large numbers of EBV particles. Chromothripsis, an integral part of this circuit has been linked to poor survival in cancers of the central nervous system, colorectal cancer, and acute myeloid leukemia (Voronina *et al.* 2020). Driver genes generated by chromoplexy (Shen, 2013) are also consistent with this model. The model does not rule out contributions from subpopulations of breast cancer cells that have differential sensitivity to cancer treatment protocols. Although they are not specific, some classical cancer drugs target centromeres and mitotic structures. These drugs may also destroy cells containing anaphase bridges so this mechanisms does not apply to every treated cancer.

The idea that bound, ubiquitous EBV sequences guide rejoining of fragments is somewhat speculative. However, this idea is consistent with a large amount of data showing EBV sequences and binding sites near breakpoints. EBV DNA gets packaged around chromatin like a small human chromosome. The viral chromosome moves around through the nuclear environment of human chromosomes, depending on the state of viral activation. In its latent form, EBV binds human repressive heterochromatin, but reactivation moves the virus to active euchromatin (Moquin *et al.* 2018). Abundant EBV binding sites exist as homologous repetitive human DNA sequences 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 produce nucleases that can also cause breaks directly (Wu *et al.* 2010), promoting telomere dysfunction and replication stress (Hafez and Luftig, 2017).

BRCA1 and BRCA2 mutations increase the numbers of breaks, but breakage-fusion-bridge cycles also occur in sporadic breast cancer chromosomes, albeit less frequently. In addition to gene defects in BRCA1 and BRCA2, other inherited gene defects 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 herpesviruses (CMV, HSV, HHV-6) are also common.

Some regions undergoing chromothripsis are largely endogenous retroviral in origin, but the sheer numbers of EBV-like sequences in human DNA rival the numbers of retrovirus homologies. This similarity reflects the constant interplay between viral- and human-DNA during evolution. Some herpes 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). Cytomegalovirus, another related herpes virus, causes congenital disabilities.

Targeting mitosis in rapidly dividing cancer cells has been effective in traditional cancer therapy. Chemotherapy agents that disrupt microtubule function have been clinically valuable and include drugs such as taxanes and vinca alkaloids. These drugs interfere with microtubules in rapidly dividing cells and alter breakage fusion bridge formation. Managing accompanying side effects, toxicity and multi drug resistance is difficult (Ruan *et al.* 2018). Strategies to prevent cancer recurrence that target breakage-fusion-bridge formation more specifically may lead to drugs that destroy cells with two centromeres.

The current evidence supports 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. S1).

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 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; Friedenson, 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)

It is unlikely that the extensive homology to EBV represents contamination of the reference human genome sequence. Most similarities occur to the unusual tumor viral variants HKHD40 and HKNPC60 and not EBV. PERV or other endogenous retroviral sequences also contribute to breast cancer breaks in many cases. Significant homology to SARS-CoV-2 also exists (Figs 5a and 5b) on

Chromosomes 2 and 12. Host factors are the primary determinant of the severity of SARS-CoV-2 infection (Zhang *et al.* 2020). Together with deregulated immune responses, these areas of homology may add genomic host factors that influence the severity of CoV-2 illness.

Sampling the population to represent the breadth of all somatic and hereditary breast cancers is not a trivial problem and limits many studies. There is no assurance that even large numbers of breast cancers are adequate 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 243 triple-negative breast cancers and 2658 human cancers (Batra *et al.* 2021; Cortes-Ciriano *et al.* 2020; Nik-Zainal *et al.* 2016; Nones *et al.* 2019; Staaf *et al.* 2019).

Materials and Methods

Characteristics of hereditary breast cancers compared to viral cancers. Selection of hereditary breast cancer genomes for this study required patient samples with a known, typed *BRCA1* or *BRCA2* gene mutation (Nik-Zainal *et al.* 2016; Nones *et al.* 2019). These hereditary cancers were mainly stage III ductal or no specific type (Nones *et al.* 2019). Seventy-four 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 without 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 whenever possible but comparisons of chromothripsis breakpoints were easier with the GrCh37/Hg19 version. 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)

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 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 human sequences and uncharacterized sample mixtures excluded. Homology among viruses was determined using "Blastn".

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 from the location of genes within EBNA1 binding sites. EBNA1 binding data was based on lymphoblastoid and naso-

pharyngeal cancer cell lines. Breaks in 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).

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. The NCBI Genome Decoration page and the Ritchie lab website (Wolfe *et al.* 2013) provided Chromosome annotation software.

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 breakpoint positions generated by random numbers. Tests for normality included kurtosis and skewness values and evaluation by methods of Shapiro-Francia and Shapiro-Wilk (Shapiro, 1965).

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; Ortutay *et al.* 2007b).

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 tested 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.

Conflict of Interest Statement

The author declares no conflicts of interest

References

Aissa AF, Islam A, Ariss MM, Go CC, Rader AE, Conrardy RD *et al* (2021). Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat Commun* **12**(1)**:** 1628.

Aleksandrov R, Hristova R, Stoynov S, Gospodinov A (2020). The Chromatin Response to Double-Strand DNA Breaks and Their Repair. *Cells* **9**(8).

Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y *et al* (2020). The repertoire of mutational signatures in human cancer. *Nature* **578**(7793)**:** 94-101.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *Journal of molecular biology* **215**(3)**:** 403-410.

Ayee R, Ofori MEO, Wright E, Quaye O (2020). Epstein Barr Virus Associated Lymphomas and Epithelia Cancers in Humans. *Journal of Cancer* **11**(7)**:** 1737-1750.

Balfour HH, Jr., Sifakis F, Sliman JA, Knight JA, Schmeling DO, 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**(8)**:** 1286-1293.

Batra RN, Lifshitz A, Vidakovic AT, Chin SF, Sati-Batra A, Sammut SJ *et al* (2021). DNA methylation landscapes of 1538 breast cancers reveal a replication-linked clock, epigenomic instability and cis-regulation. *Nat Commun* **12**(1)**:** 5406.

Bhela S, Mulik S, Reddy PB, Richardson RL, Gimenez F, Rajasagi NK *et al* (2014). Critical role of microRNA-155 in herpes simplex encephalitis. *Journal of immunology* **192**(6)**:** 2734-2743.

Bobrovnitchaia I, Valieris R, Drummond RD, Lima JP, Freitas HC, Bartelli TF *et al* (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**(1)**:** 181-191.

Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R *et al* (2013). InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation. *Nucleic acids research* **41**(Database issue)**:** D1228-1233.

Bruce JP, To KF, Lui VWY, Chung GTY, Chan YY, Tsang CM *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**(1)**:** 4193.

Bruce JP, Yip K, Bratman SV, Ito E, Liu FF (2015). Nasopharyngeal Cancer: Molecular Landscape. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **33**(29)**:** 3346-3355.

Busch K, Keller T, Fuchs U, Yeh RF, Harbott J, Klose I *et al* (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**(8)**:** 1739-1751.

Buschle A, Mrozek-Gorska P, Cernilogar FM, Ettinger A, Pich D, Krebs S *et al* (2021). Epstein-Barr virus inactivates the transcriptome and disrupts the chromatin architecture of its host cell in the first phase of lytic reactivation. *Nucleic acids research* **49**(6)**:** 3217-3241.

Charoentong P, Finotello F, Angelova M, Mayer C, Efremova M, Rieder D *et al* (2017). Pan-cancer Immunogenomic Analyses Reveal Genotype-Immunophenotype Relationships and Predictors of Response to Checkpoint Blockade. *Cell reports* **18**(1)**:** 248-262.

Chin K, de Solorzano CO, Knowles D, Jones A, Chou W, Rodriguez EG *et al* (2004). In situ analyses of genome instability in breast cancer. *Nature genetics* **36**(9)**:** 984-988.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A *et al* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**(11)**:** 1422-1423.

Cortes-Ciriano I, Lee JJ, Xi R, Jain D, Jung YL, Yang L *et al* (2020). Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nature genetics* **52**(3)**:** 331-341.

Cuceu C, Hempel WM, Sabatier L, Bosq J, Carde P, M'Kacher R (2018). Chromosomal Instability in Hodgkin Lymphoma: An In-Depth Review and Perspectives. *Cancers (Basel)* **10**(4).

Denner J (2017). The porcine virome and xenotransplantation. *Virology journal* **14**(1)**:** 171.

Fina F, Romain S, Ouafik L, Palmari J, Ben Ayed F, Benharkat S *et al* (2001). Frequency and genome load of Epstein-Barr virus in 509 breast cancers from different geographical areas. *British journal of cancer* **84**(6)**:** 783-790.

Friedenson B (2009). Dewey defeats Truman and cancer statistics. *Journal of the National Cancer Institute* **101**(16)**:** 1157.

Friedenson B (2013). Mutations in components of antiviral or microbial defense as a basis for breast cancer. *Funct Integr Genomics* **13**(4)**:** 411-424.

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.

Funk CC, Casella AM, Jung S, Richards MA, Rodriguez A, Shannon P *et al* (2020). Atlas of Transcription Factor Binding Sites from ENCODE DNase Hypersensitivity Data across 27 Tissue Types. *Cell reports* **32**(7)**:** 108029.

Germini D, Sall FB, Shmakova A, Wiels J, Dokudovskaya S, Drouet E *et al* (2020). Oncogenic Properties of the EBV ZEBRA Protein. *Cancers (Basel)* **12**(6).

Hafez AY, Luftig MA (2017). Characterization of the EBV-Induced Persistent DNA Damage Response. *Viruses* **9**(12).

Hau PM, Lung HL, Wu M, Tsang CM, Wong KL, Mak NK *et al* (2020). Targeting Epstein-Barr Virus in Nasopharyngeal Carcinoma. *Front Oncol* **10:** 600.

Helman E, Lawrence MS, Stewart C, Sougnez C, Getz G, Meyerson M (2014). Somatic retrotransposition in human cancer revealed by whole-genome and exome sequencing. *Genome research* **24**(7)**:** 1053-1063.

Hu H, Luo ML, Desmedt C, Nabavi S, Yadegarynia S, Hong A *et al* (2016). Epstein-Barr Virus Infection of Mammary Epithelial Cells Promotes Malignant Transformation. *EBioMedicine* **9:** 148-160.

Kamranvar SA, Gruhne B, Szeles A, Masucci MG (2007). Epstein-Barr virus promotes genomic instability in Burkitt's lymphoma. *Oncogene* **26**(35)**:** 5115-5123.

Kim KD, Tanizawa H, De Leo A, Vladimirova O, Kossenkov A, Lu F *et al* (2020). Epigenetic specifications of host chromosome docking sites for latent Epstein-Barr virus. *Nat Commun* **11**(1)**:** 877.

Kioussis D, Pachnis V (2009). Immune and nervous systems: more than just a superficial similarity? *Immunity* **31**(5)**:** 705-710.

Kumar R, Nagpal G, Kumar V, Usmani SS, Agrawal P, Raghava GPS (2019). HumCFS: a database of fragile sites in human chromosomes. *BMC genomics* **19**(Suppl 9)**:** 985.

Lacoste S, Wiechec E, Dos Santos Silva AG, Guffei A, Williams G, Lowbeer M *et al* (2010). Chromosomal rearrangements after ex vivo Epstein-Barr virus (EBV) infection of human B cells. *Oncogene* **29**(4)**:** 503-515.

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

Law EK, Levin-Klein R, Jarvis MC, Kim H, Argyris PP, Carpenter MA *et al* (2020). APOBEC3A catalyzes mutation and drives carcinogenesis in vivo. *The Journal of experimental medicine* **217**(12).

Lawson JS, Glenn WK (2021). Catching viral breast cancer. *Infectious agents and cancer* **16**(1)**:** 37.

Lee Y, Kang JK, Lee YH, Yoon HJ, Yang SS, Kim SH *et al* (2021). Chromosome aberration dynamics in breast cancer patients treated with radiotherapy: Implications for radiation biodosimetry. *Mutat Res Genet Toxicol Environ Mutagen* **872:** 503419.

Leibowitz ML, Zhang CZ, Pellman D (2015). Chromothripsis: A New Mechanism for Rapid Karyotype Evolution. *Annu Rev Genet* **49:** 183-211.

Li WF, Zhang Y, Huang XB, Du XJ, Tang LL, Chen L *et al* (2017). Prognostic value of plasma Epstein-Barr virus DNA level during posttreatment follow-up in the patients with nasopharyngeal carcinoma having undergone intensity-modulated radiotherapy. *Chin J Cancer* **36**(1)**:** 87.

Li Y, Roberts ND, Wala JA, Shapira O, Schumacher SE, Kumar K *et al* (2020). Patterns of somatic structural variation in human cancer genomes. *Nature* **578**(7793)**:** 112-121.

Lopez C, Kleinheinz K, Aukema SM, Rohde M, Bern-

hart SH, Hubschmann D *et al* (2019). Genomic and transcriptomic changes complement each other in the pathogenesis of sporadic Burkitt lymphoma. *Nat Commun* **10**(1)**:** 1459.

Lorenzetti MA, De Matteo E, Gass H, Martinez Vazquez P, Lara J, Gonzalez P *et al* (2010). Characterization of Epstein Barr virus latency pattern in Argentine breast carcinoma. *PloS one* **5**(10)**:** e13603.

Lu F, Wikramasinghe P, Norseen J, Tsai K, Wang P, Showe L *et al* (2010). Genome-wide analysis of host-chromosome binding sites for Epstein-Barr Virus Nuclear Antigen 1 (EBNA1). *Virology journal* **7:** 262.

Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A *et al* (2008). InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Molecular systems biology* **4:** 218.

Maccaroni K, Balzano E, Mirimao F, Giunta S, Pelliccia F (2020). Impaired Replication Timing Promotes Tissue-Specific Expression of Common Fragile Sites. *Genes (Basel)* **11**(3).

Maciejowski J, Chatzipli A, Dananberg A, Chu K, Toufektchan E, Klimczak LJ *et al* (2020). APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. *Nature genetics* **52**(9)**:** 884-890.

Mameli G, Poddighe L, Mei A, Uleri E, Sotgiu S, Serra C *et al* (2012). Expression and activation by Epstein Barr virus of human endogenous retroviruses-W in blood cells and astrocytes: inference for multiple sclerosis. *PloS one* **7**(9)**:** e44991.

Marrao G, Habib M, Paiva A, Bicout D, Fallecker C, Franco S *et al* (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.

McClintock B (1941). The stability of broken ends of chromosomes in Zea Mays. *Genetics* **26:** 234-282.

Meier UC, Cipian RC, Karimi A, Ramasamy R, Middeldorp JM (2021). Cumulative Roles for Epstein-Barr Virus, Human Endogenous Retroviruses, and Human Herpes Virus-6 in Driving an Inflammatory Cascade Underlying MS Pathogenesis. *Frontiers in immunology* **12:** 757302.

Moquin SA, Thomas S, Whalen S, Warburton A, Fernandez SG, McBride AA *et al* (2018). The Epstein-Barr Virus Episome Maneuvers between Nuclear Chromatin Compartments during Reactivation. *Journal of virology* **92**(3).

Mount DW (2007). Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc* **2007:** pdb top17.

Nakshatri H, Bhat-Nakshatri P, Martin DA, Goulet RJ, Jr., Sledge GW, Jr. (1997). Constitutive activation of NF-kappaB during progression of breast cancer to hormone-independent growth. *Molecular and cellular biology* **17**(7)**:** 3629-3639.

Nik-Zainal S, Davies H, Staaf J, Ramakrishna M, Glodzik D, Zou X *et al* (2016). Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**(7605)**:** 47-54.

Nik-Zainal S, 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**(11)**:** 2617-2629.

Nones K, Johnson J, Newell F, Patch AM, Thorne H, Kazakoff SH *et al* (2019). Whole-genome sequencing reveals clinically relevant insights into the aetiology of familial breast cancers. *Ann Oncol* **30**(7)**:** 1071-1079.

Ortutay C, Siermala M, Vihinen M (2007a). ImmTree: database of evolutionary relationships of genes and proteins in the human immune system. *Immunome Res* **3:** 4.

Ortutay C, Siermala M, Vihinen M (2007b). Molecular characterization of the immune system: emergence of proteins, processes, and domains. *Immunogenetics* **59**(5)**:** 333-348.

Ou X, Shen, C, He, X, Hu C (2020). Germline Mutation Landscape of DNA Damage Repair Genes in Nasopharyngeal Carcinoma. *International Journal of Radiation Oncology • Biology • Physics* **108:** S50.

Pan H, Gray R, Braybrooke J, Davies C, Taylor C, McGale P *et al* (2017). 20-Year Risks of Breast-Cancer Recurrence after Stopping Endocrine Therapy at 5 Years. *The New England journal of medicine* **377**(19)**:** 1836-1846.

Prusty BK, zur Hausen H, Schmidt R, Kimmel R, de Villiers EM (2008). Transcription of HERV-E and HERV-E-related sequences in malignant and non-malignant human haematopoietic cells. *Virology* **382**(1)**:** 37-45.

Ruan W, Lim HH, Surana U (2018). Mapping Mitotic Death: Functional Integration of Mitochondria, Spindle Assembly Checkpoint and Apoptosis. *Front Cell Dev Biol* **6:** 177.

Sarid R, Gao SJ (2011). Viruses and human cancer: from detection to causality. *Cancer letters* **305**(2)**:** 218-227.

Scott RS (2017). Epstein-Barr virus: a master epigenetic manipulator. *Curr Opin Virol* **26:** 74-80.

Shapiro RL (1965). An analysis of variance test for normality (complete samples). *Biometrika* **52**(3-4)**:** 20.

Shen MM (2013). Chromoplexy: a new category of complex rearrangements in the cancer genome. *Cancer cell* **23**(5)**:** 567-569.

Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reutersward C, Hakkinen J *et al* (2019). Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nature medicine* **25**(10)**:** 1526-1533.

Stephens PJ, Greenman CD, Fu B, Yang F, Bignell GR, Mudie LJ *et al* (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**(1)**:** 27-40.

Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC *et al* (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**(7403)**:** 400-404.

Tang MH, Varadan V, Kamalakaran S, Zhang MQ, Dimitrova N, Hicks J (2012). Major chromosomal breakpoint intervals in breast cancer co-localize with differentially methylated regions. *Front Oncol* **2:** 197.

Tsurumi T, Kobayashi A, Tamai K, Yamada H, Daikoku T, Yamashita Y *et al* (1996). Epstein-Barr virus single-stranded DNA-binding protein: purification, characterization, and action on DNA synthesis by the viral DNA polymerase. *Virology* **222**(2)**:** 352-364.

Tweedy J, Spyrou MA, Pearson M, Lassner D, Kuhl U, Gompels UA (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**(1).

Umbreit NT, Zhang CZ, Lynch LD, Blaine LJ, Cheng AM, Tourdot R *et al* (2020). Mechanisms generating cancer genome complexity from a single cell division error. *Science* **368**(6488).

Voronina N, Wong JKL, Hubschmann D, Hlevnjak M, Uhrig S, Heilig CE *et al* (2020). The landscape of chromothripsis across adult cancer types. *Nat Commun* **11**(1)**:** 2320.

Wolfe D, Dudek S, Ritchie MD, Pendergrass SA (2013). Visualizing genomic information across chromosomes with PhenoGram. *BioData Min* **6**(1)**:** 18.

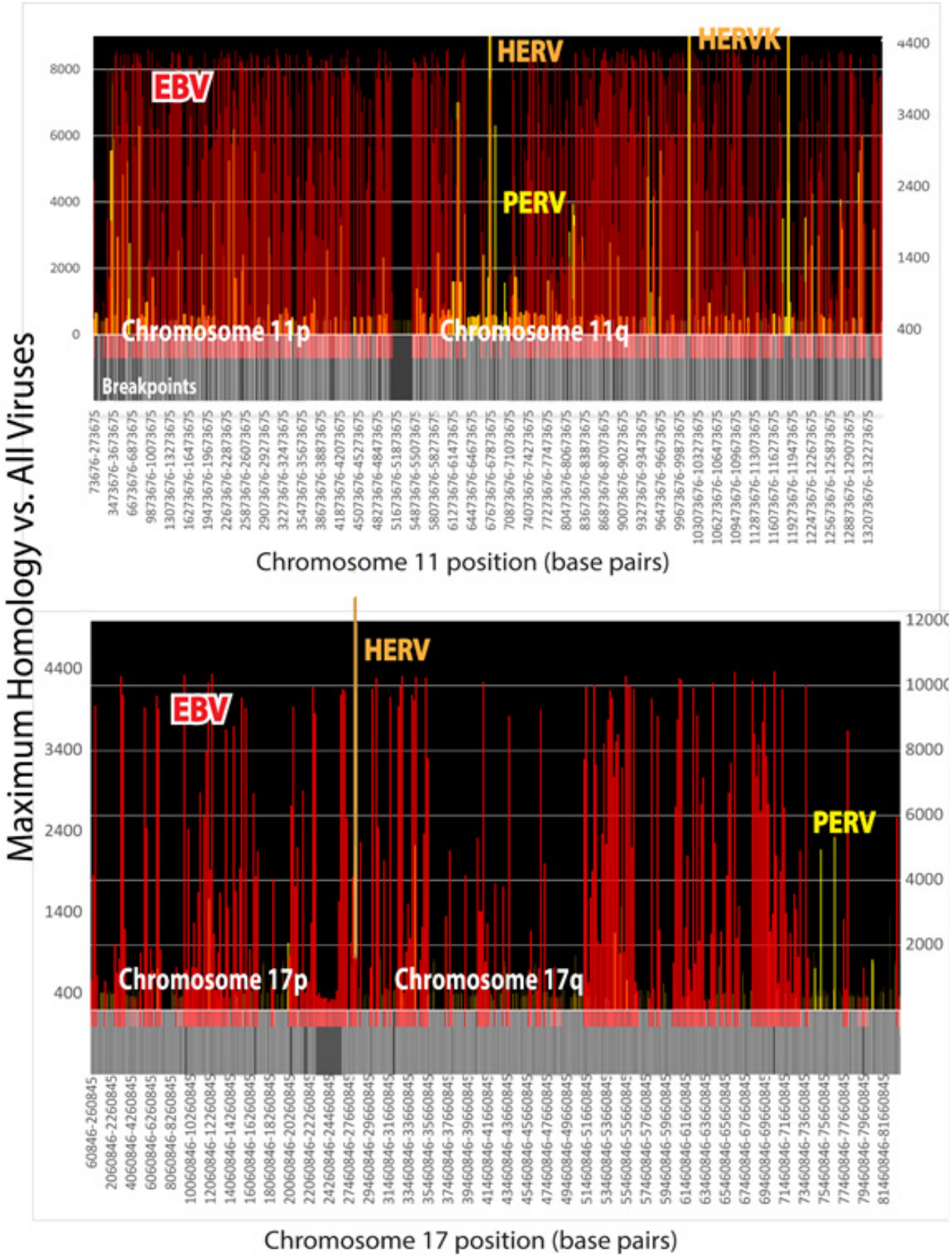
Wu CC, Liu MT, Chang YT, Fang CY, Chou SP, Liao HW *et al* (2010). Epstein-Barr virus DNase (BGLF5) induces genomic instability in human epithelial cells. *Nucleic acids research* **38**(6)**:** 1932-1949.

Xiao K, Yu Z, Li X, Li X, Tang K, Tu C *et al* (2016). Genome-wide Analysis of Epstein-Barr Virus (EBV) Integration and Strain in C666-1 and Raji Cells. *Journal of Cancer* **7**(2)**:** 214-224.

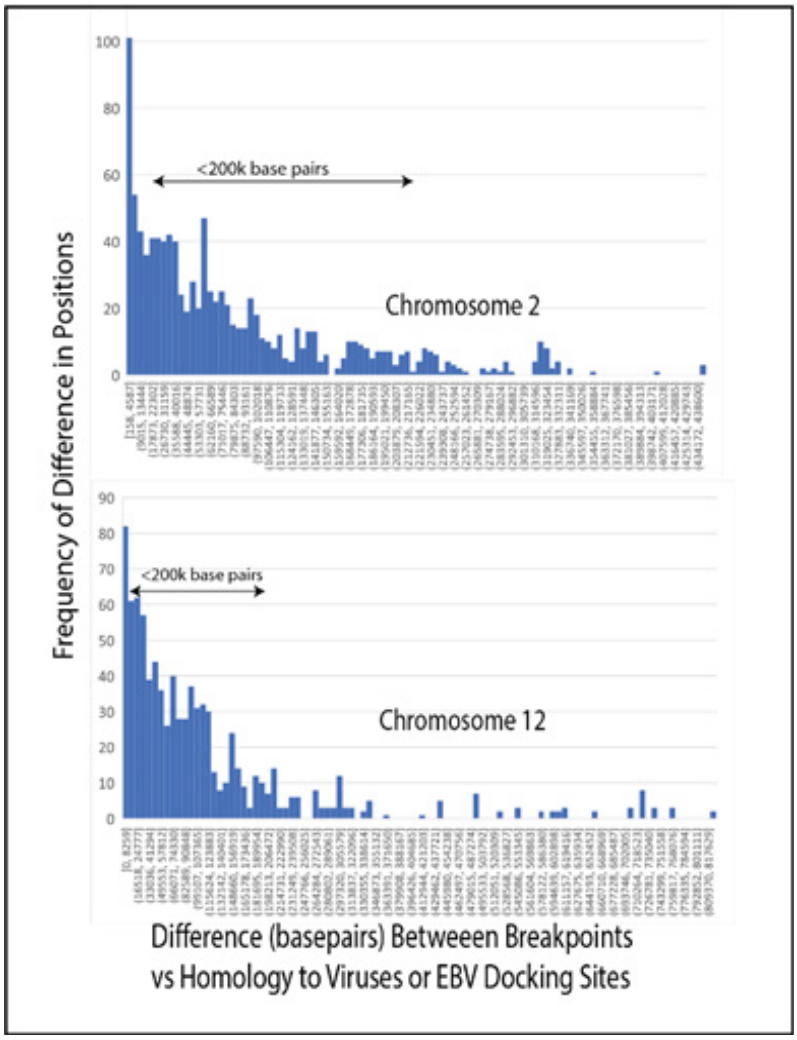
Xu M, Yao Y, Chen H, Zhang S, Cao SM, Zhang Z *et al* (2019). Genome sequencing analysis identifies Epstein-Barr virus subtypes associated with high risk of nasopharyngeal carcinoma. *Nature genetics* **51**(7)**:** 1131-1136.

Zhang X, Tan Y, Ling Y, Lu G, Liu F, Yi Z *et al* (2020). Viral and host factors related to the clinical outcome of COVID-19. *Nature* **583**(7816)**:** 437-440.

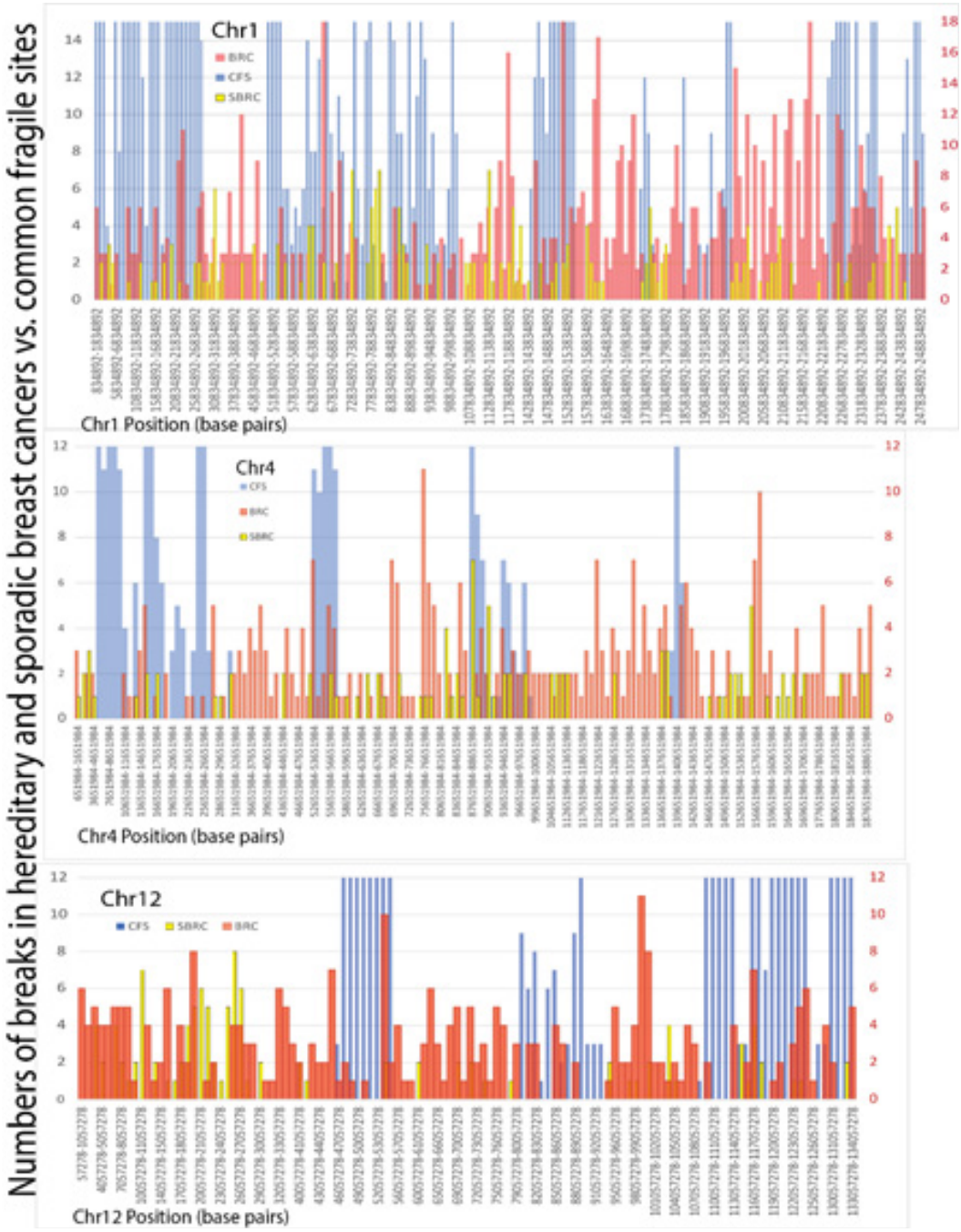
Zhang Z, Schwartz S, Wagner L, Miller W (2000). A greedy algorithm for aligning DNA sequences. *J Comput Biol* **7**(1-2)**:** 203-214.



Supplementary Fig. S1. 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).



Supplementary Fig. S2. Distribution of breakpoint chromosome positions in 139 hereditary or likely hereditary breast cancers on chromosome 2 and chromosome 12 relative to areas of homology to viruses or EBV docking sites



.

Supplementary Fig S3. 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.