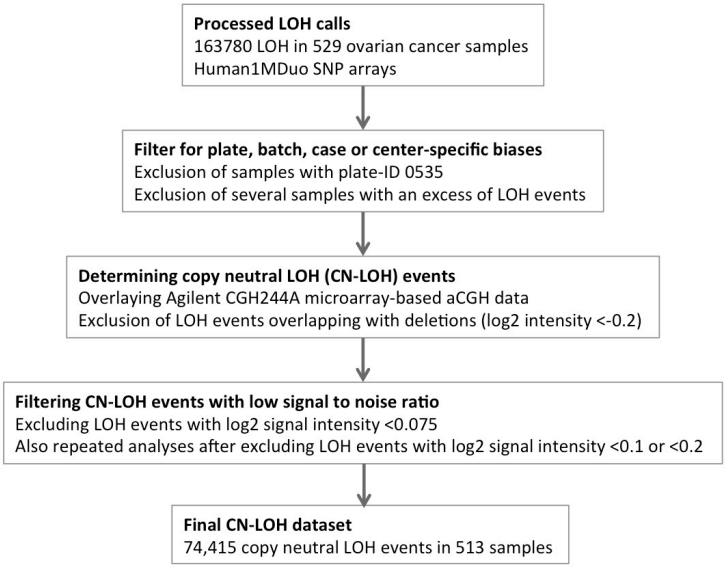
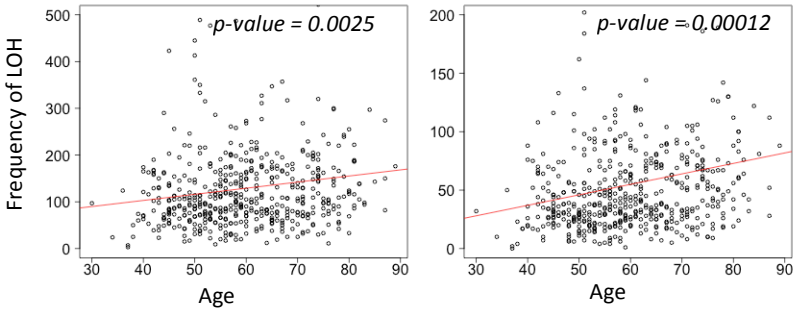
Supplementary Figure SF1



Supplementary Figure SF1: A schematic diagram showing the processing of LOH dataset.

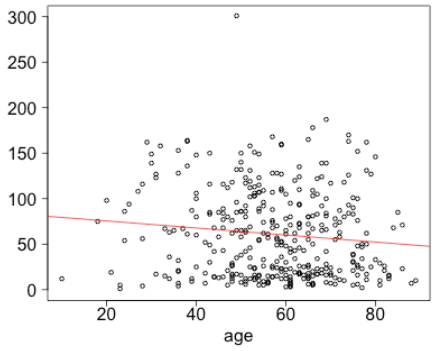
Supplementary Figure SF2



Supplementary Figure SF2: Association between the frequency of LOH events and age for different thresholds for LOH log2 signal intensity (left) 0.1 and (right) 0.2. The trend-line is shown in red. P-value of significance of the association between age and the LOH frequency is shown at the top-right corner of each panel.

Supplementary text Stx1 and Supplementary Figure SF3

We obtained data on loss of heterozygosity (LOH), somatic copy number, and clinical information for 357 glioblastoma samples from the Cancer Genome Atlas (TCGA (2008)). LOH and copy number status were determined using Illumina Human1MDuo SNPchip (at the Hudson Alpha Institute for Biotechnology) and HG-CGH-244A arrays (at Harvard Medical School), respectively, as a part of the TCGA initiative. Combining the LOH and copy number calls, we identified the copy neutral LOH events in the glioblastoma samples, in a manner similar to that for the ovarian cancer samples, as described in the Methods section of the main text. Overlaying data for 'age at initial diagnosis', we found that there was a weak and negative correlation (p-value >0.01) between age and the number of LOH events (**Supplementary Figure SF3**). Moreover, we did not find any preference for frequent LOH of chr17. These results could be biologically relevant indicating distinct biology of glioblastoma. But previous studies have pointed out increased noise and systematic biases in the TCGA data (Leek, et al. 2010), and so, we prefer to interpret the glioblastoma results cautiously.



Supplementary Figure SF3: Scatter plot showing the weak and negative correlation (p-value: 0.03) between the frequency of LOH events in the glioblastoma patients against their age at initial diagnosis. The red line indicates the trend line of the linear regression.

Supplementary Figure SF4

```
lm(formula = LOH ~ Age + BRCA1 + BRCA2 + RAD50 + RAD51 + RAD52 +
    RAD54B + RAD54L + XRCC2 + XRCC3 + MRE11A, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-222.70  -63.60  -21.19   29.48 1462.29

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -168.1574   164.1591  -1.024  0.30627
Age            1.4686     0.5501   2.670  0.00789 **
BRCA1         -9.8750    15.8297  -0.624  0.53309
BRCA2          5.2102    22.3800   0.233  0.81603
RAD50         -0.8255    11.4358  -0.072  0.94249
RAD51         22.3340    19.9841   1.118  0.26440
RAD52         -8.0953    21.0704  -0.384  0.70103
RAD54B       -19.9276    13.8582  -1.438  0.15121
RAD54L        35.5702    13.7986   2.578  0.01029 *
XRCC2          9.3951    11.9403   0.787  0.43183
XRCC3         10.3735    22.9934   0.451  0.65212
MRE11A         2.6225    19.4766   0.135  0.89296
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 120.5 on 408 degrees of freedom
Multiple R-squared: 0.07111,     Adjusted R-squared: 0.04607
F-statistic: 2.839 on 11 and 408 DF,  p-value: 0.001358
```

Supplementary Figure SF4: Coefficients in the multivariate model testing for association between the frequency of LOH events with age and expression levels of several key genes in the homologous recombination pathway.

Supplementary Text STx2

To estimate pathway-level imbalance using an alternative parametric approach, we first analyzed expression data of a set of genes (e.g. HR pathway genes (Chapman, et al. 2012)) in the tumor samples alongside that in one (or average) normal sample. For each gene (*i*) in each sample (*j*), we calculated the normalized difference in its expression level in the sample with that in the normal control (d_{ij} = $Expr_i^{tumor\text{-}j}$ – $Expr_i^{normal}$/ $Expr_i^{normal}$). If a sample has relatively high expression (compared to control) for all the genes (*i* in 1:N) in the set, average $d_{ij}^{i=1:N}$ will be high and variance (e.g. stdev($d_{ij}^{i=1:N}$)) would be small. Similarly, for a sample with relatively low (or moderate) expression for all the genes in the set, will have low (or intermediate) average $d_{ij}^{i=1:N}$ and small variance. In contrast, a sample with relatively high expression of some genes (high d_{ij}), and low expression in some other (low d_{ij}) will have a high stdev($d_{ij}^{i=1:N}$). We tested whether stdev($d_{ij}^{i=1:N}$) significantly correlate with the number of LOH events per sample, and found that it is also significantly correlated (p-value: 0.000941). Furthermore, consistent with our previous results (Figure 3), the effect of this parametric measure of pathway level imbalance and age have significant effects on age, which exist even after adjusting for one another (p-value <0.01 in both cases). However, the parametric measure was sensitive to outliers, and thus we preferred the non-parametric score used in Figure 3 in the main text.

Supplementary References

2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216):1061-1068.

Chapman JR, Taylor MR, Boulton SJ. 2012. Playing the end game: DNA double-strand break repair pathway choice. Mol Cell 47(4):497-510.

Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11(10):733-739.