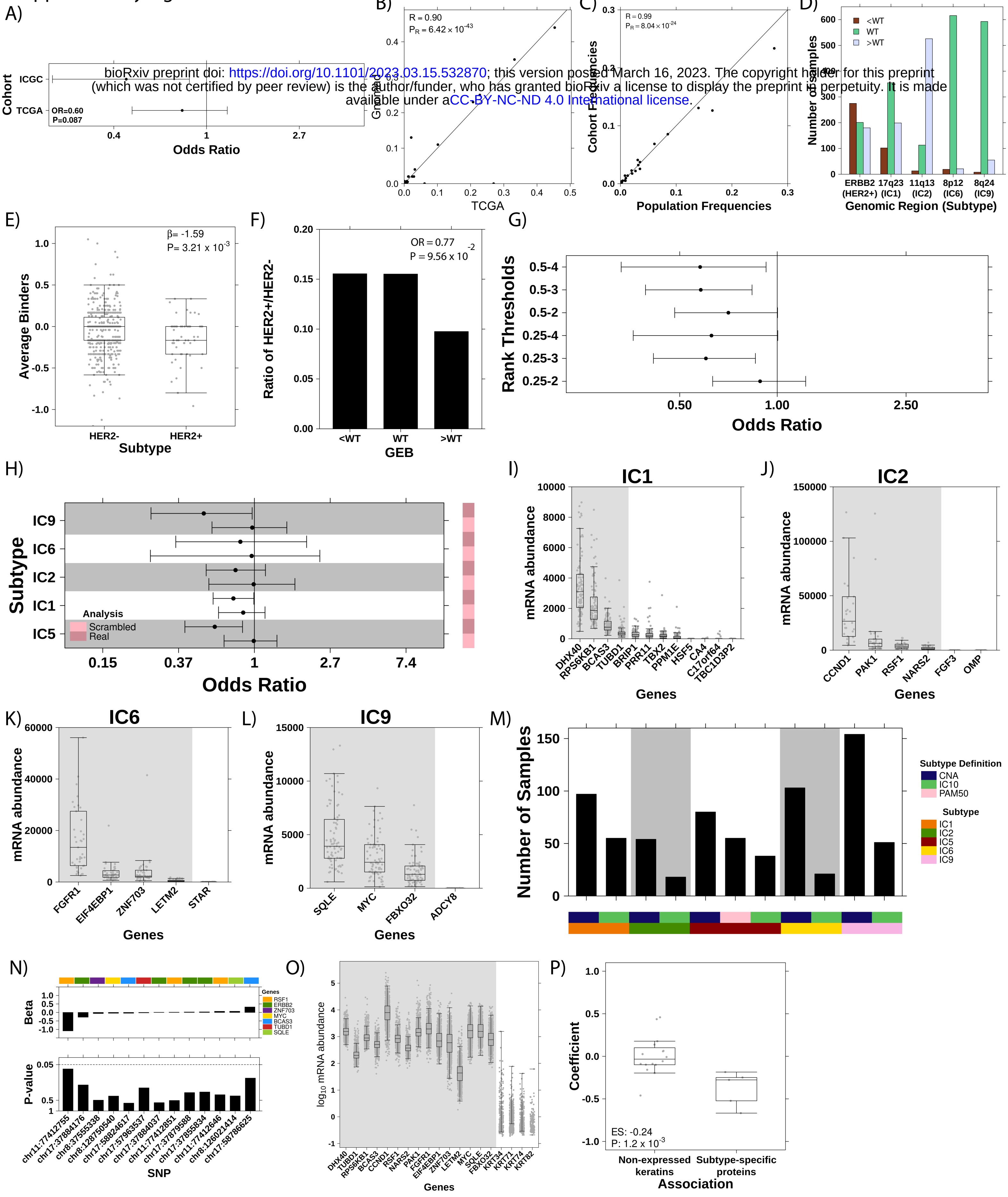
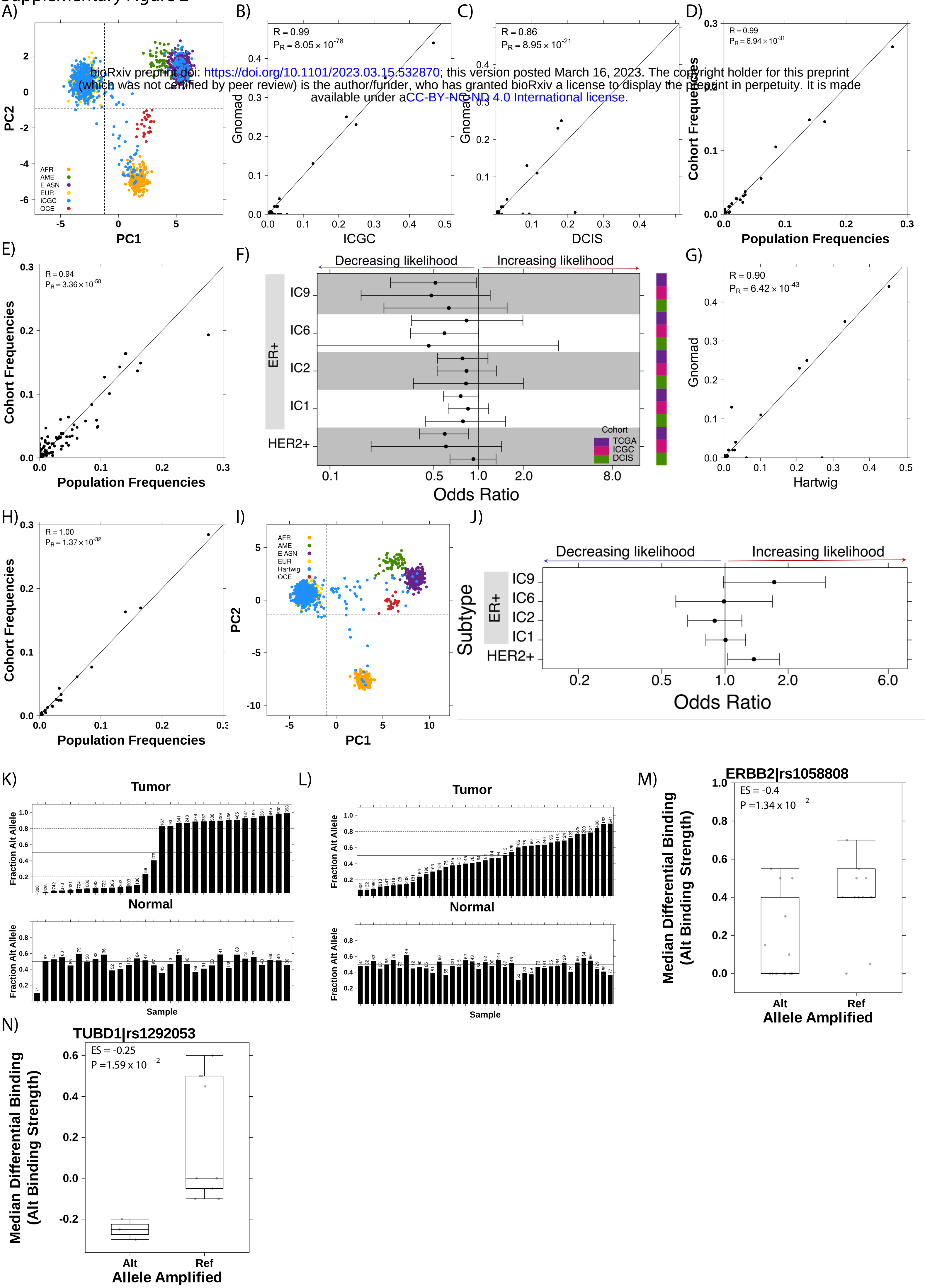
Supplementary Figure 1



Supplementary Figure 1 – GEB in oncogene selects against oncogene amplification
A) Forest plot shows OR and 95% confidence intervals for association between number of HLA alleles an individual possesses that can bind GP2 and whether the individual has HER2+ breast cancer in ICGC and TCGA. **B)** Scatterplot of minor allele frequencies (MAF) in TCGA discovery cohort compared to population frequencies in Gnomad (Pearson correlation). **C)** Scatterplot of HLA allele frequencies in TCGA discovery cohort compared to population frequencies in The Allele Frequency Net Database. **D)** Barplot showing the number of samples that have low, medium or high GEB (defined as less than the reference genome (<WT), the same as reference (WT) or greater than the reference genome (>WT)) in subtype specific recurrently amplified loci. **E)** Boxplot showing depletion of the average number of binders in *ERBB2* in HER2+ breast cancer compared to HER2- breast cancer. Statistics from a logistic regression model correcting for the first six genetic principal components. Boxplot represents median, 0.25 and 0.75 quantiles with whickers at 1.5x interquartile range. **F)** Barplot shows the ratio of HER2+ to HER2- patients with low, medium or high GEB defining HER2+ as having an *ERBB2* amplification (*i.e.* >4 copies). Statistics from logistic regression model correcting for the first six genetic principal components. **G)** Scatterplot showing odds ratio (x-axis) between GEB and HER2+ breast cancer considering varying definition of HLA binders (y-axis). **H)** Negative association between GEB and subtype commitment is not driven by germline variants alone. Forest plot shows odds ratio and 95% confidence intervals for the true associations ("real") compared to associations run with scrambled HLA alleles ("scrambled"). Odds ratio and 95% confidence interval plotted were calculated as the median, 0.025 and 0.975 quantiles of 1,000 iterations of scrambled HLA alleles. Covariate along the right indicates if statistics are from real or scrambled analyses. **I-L)** mRNA abundance of recurrently amplified genes in each of the four high risk ER+ IntClust subtypes: IC1 **(I)**, IC2 **(J)**, IC6 **(K)** and IC9 **(L)**. **M)** Boxplot shows the number of samples (y-axis) corresponding to each subtype (x-axis) based on alternative subtype definitions. **N)** Barplot shows effect size (top) and p-value (bottom) from association with breast cancer risk from Zhang *et al.* (*7*) **O-P)** As a negative control, we tested the association of the GEB in unexpressed keratins, KRT34, KRT71, KRT74 and KRT82, with the PAM50 subtypes. As these proteins are not expressed in mammary tissue, there should be no association. **(O)** Boxplot shows $\log_{10}$ mRNA abundance of subtype specific genes on the left (grey background shading) compared to the unexpressed keratins on the right. **(P)** Boxplot shows the coefficients for these analyses are significantly closer to zero compared to the coefficients from the subtype-specific protein analyses. Effect size (ES) represents the difference in medians. P-value from Mann Whitney Rank Sum Test.

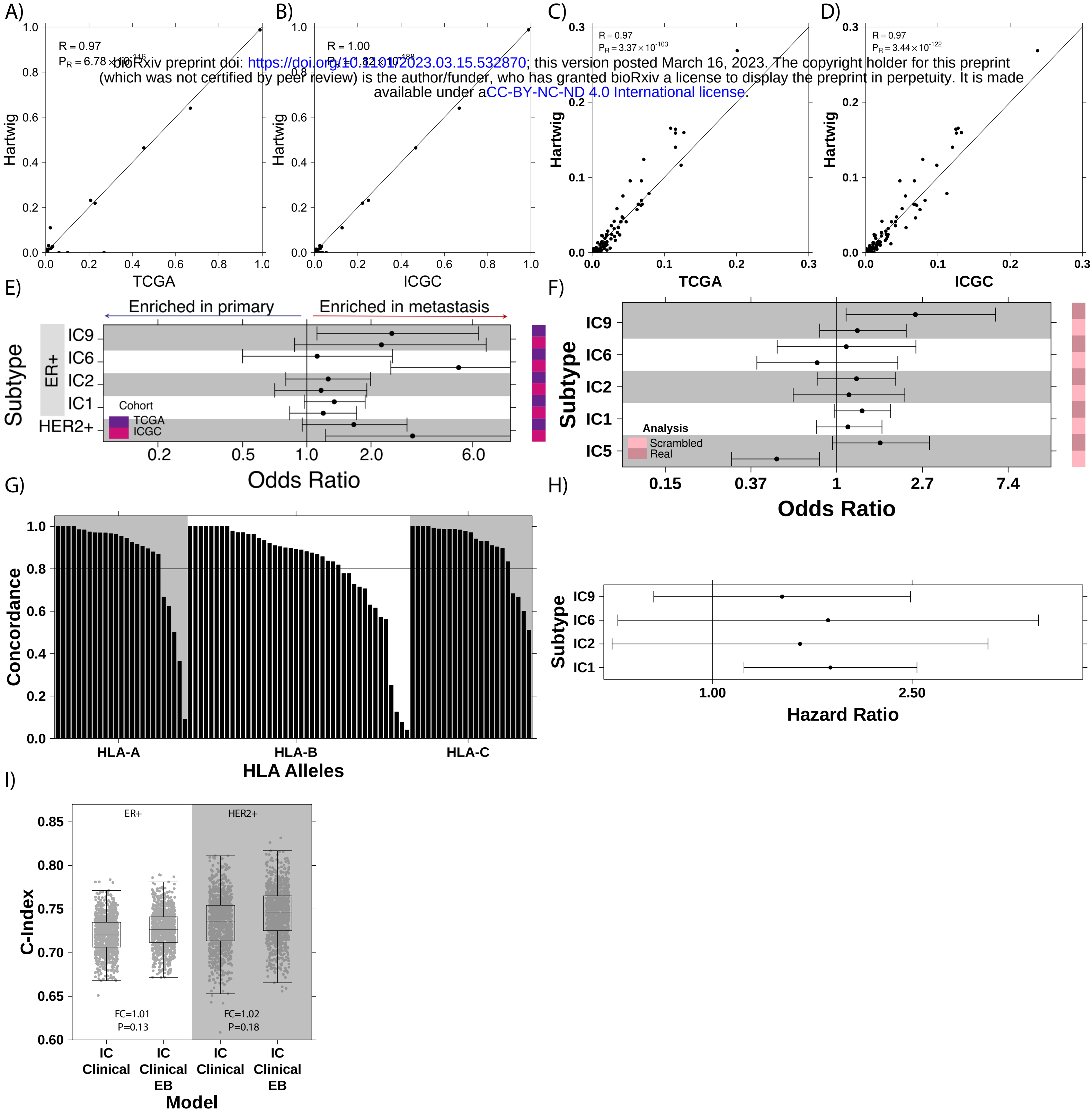
Supplementary Figure 2



Supplementary Figure 2 – Germline-mediated immunoediting dictates breast cancer subtype early during tumorigenesis

A) Scatterplot shows principal component 1 and 2 for ICGC replication cohort against reference cohort based on 128 ancestry informative markers (*51*). Majority of samples cluster with European population. Dotted lines represent cutoffs for inclusion in analysis. **B-C)** Scatterplot of minor allele frequencies (MAF) in ICGC **(B)** and DCIS **(C)** replication cohorts compared to population frequencies in Gnomad (Pearson correlation). **D-E)** Scatterplot of HLA allele frequencies in ICGC **(D)** and DCIS **(E)** replication cohorts compared to population frequencies in The Allele Frequency Net Database. **F)** Across five subtypes and three independent cohorts, a high GEB in subtype-specific oncogenes is associated with decreased likelihood of developing the cognate subtype. Forest plot shows the odds ratio and 95% confidence intervals for three cohorts: DCIS, TCGA and ICGC. Covariate on the right indicates cohort. **G)** Scatterplot of minor allele frequencies (MAF) Hartwig cohort compared to population frequencies in Gnomad (Pearson correlation). **H)** Scatterplot of HLA allele frequencies in Hartwig cohort compared to population frequencies in The Allele Frequency Net Database. **I)** Scatterplot shows principal component 1 and 2 for Hartwig cohort against reference cohort based on 128 ancestry informative markers (*51*). Majority of samples cluster with European population. Dotted lines represent cutoffs for inclusion in analysis. **J)** Forest plot shows association between GEB and subtype in metastatic breast cancer (Hartwig). No association was observed in metastatic breast cancer. **K-L)** Barplots show fraction of reads supporting the alternative allele in the tumor (top) and the normal (bottom) for two common variants: rs1058808 **(K)** and rs1292053 (**L**). The number on the top of each plot shows the top number of reads covering each loci. The horizontal line indicates fraction = 0.5 while the dotted lines represent fraction = 0.2 or 0.8. **M-N)** Boxplots of median differential binding per sample for epitopes derived from the alt allele *vs* ref allele (y-axis) for samples that preferentially amplified the alt or the ref allele. Effect size and p-value from Mann-Whitney rank sum test. Boxplots show analysis for rs1058808 derived from *ERBB2* **(M)** and rs1292053 derived from *TUBD1* **(N)**.

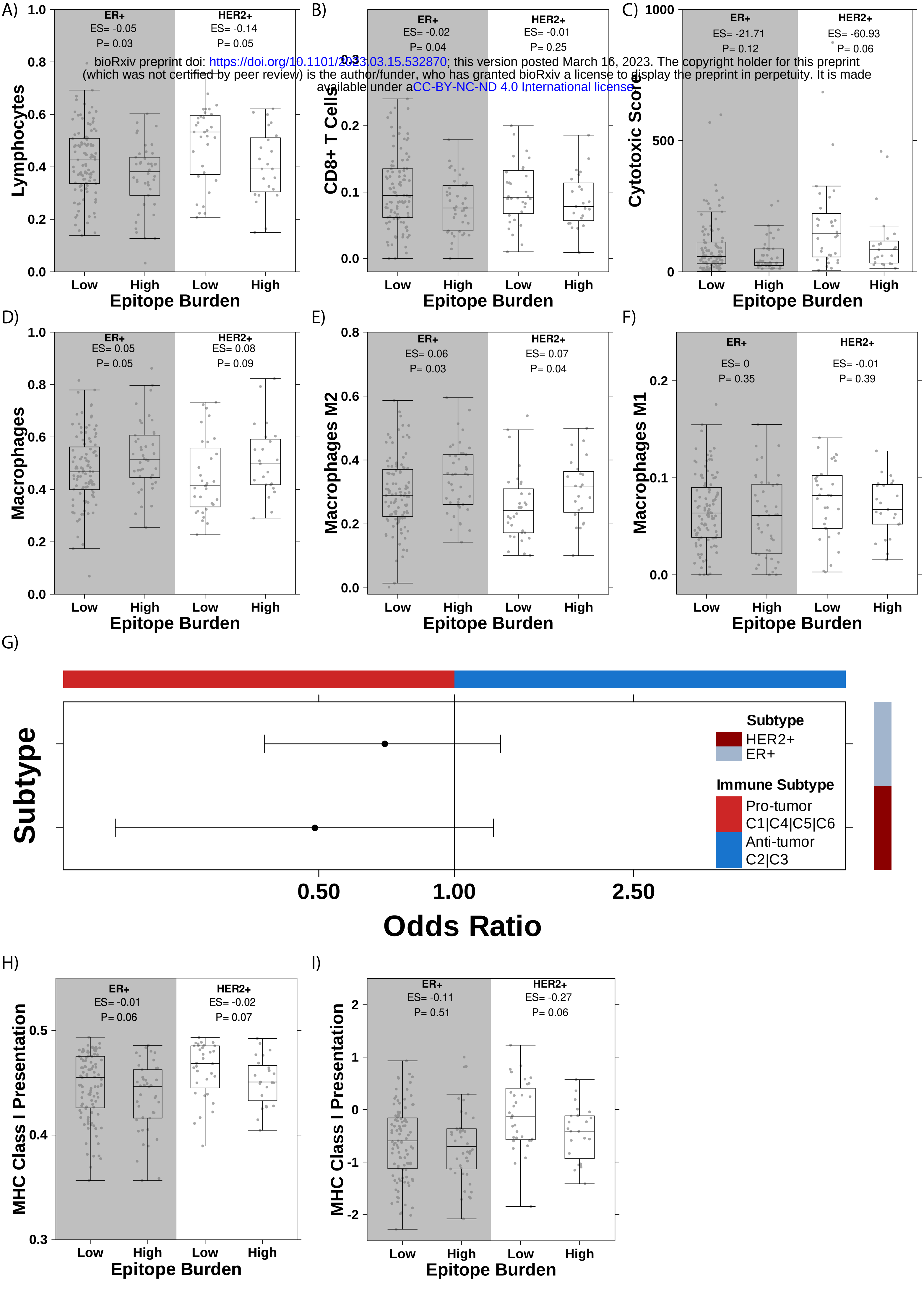
Supplementary Figure 3



Supplementary Figure 3 – Tumors that overcome a high GEB are more aggressive

A-B) Scatterplot of minor allele frequencies (MAF) in TCGA **(A)** and ICGC **(B)** compared to Hartwig (Pearson correlation). **C-D)** Scatterplot of HLA allele frequencies in TCGA **(C)** and ICGC **(D)** compared to Hartwig (Pearson correlation). **E)** Forest plot comparing GEB between primary and metastatic tumors of the same subtype. Two primary cohorts were evaluated, TCGA and ICGC, against one metastatic cohort (Hartwig). Covariate on the right indicate which cohort. **F)** Enrichment in metastatic tumors is not driven by germline variants alone. Forest plot shows odds ratio and 95% confidence intervals for the true associations ("real") compared to associations run with scrambled HLA alleles ("scrambled"). Odds ratio and 95% confidence interval plotted were calculated as the median, 0.025 and 0.975 quantiles of 1,000 iterations of scrambled HLA alleles. Covariate along the right indicates if statistics are from real or scrambled analyses. **G)** Accuracy of HLA imputation from TCGA SNP6 data compared to HLA genotyping from WES from Polysolver (*49*) as the gold standard. Horizontal line indicates accuracy of 80%. **H)** Forest plot shows hazard ratio (HR) and 95% confidence intervals from CoxPH correcting for the first two genetic principal components, age and percent genome altered (PGA) for each high-risk ER+ subtype individually. **I)** Boxplot shows c-index of predictive models considering Integrative Clusters (IC) and clinicopathologic features (age, size, grade and node involvement) alone or in combination with GEB for 1,000 bootstrapped iterations. Fold change (FC) is calculated as the ratio of medians while the p-value is calculated as 1 – the proportion of iterations where the c-index of the IC+clinical+GEB model was greater than the IC+clinical model.

Supplementary Figure 4



Supplementary Figure 4 – A high GEB promotes an immunosuppressive phenotype
A-F) Lymphocyte infiltration **(A)**, CD8+ T cells infiltration **(B)**, cytotoxic score **(C)**, macrophage infiltration **(D)**, M2- **(E)** or M1-polarized macrophages **(F)** in ER+ or HER2+ high germline epitope tumors compared to low germline epitope tumors in TCGA. Effect size (ES) shows difference in medians while p-value is from Mann-Whitney Rank Sum test. Boxplot represents median, 0.25 and 0.75 quantiles with whickers at 1.5x interquartile range. **g)** Forest plot shows odds of developing anti-tumor immune subtype (x-axis) given a high GEB in HER2+ or ER+ subtypes (y-axis). Covariate along the right indicate the subtype evaluated while the covariate along the top indicates the interpretation of the direction of effect. **H-I)** Boxplot of MHC Class I antigen presentation pathway measured by two different transcriptional signatures (y-axis) stratified by high *vs* low GEB tumors (x-axis). Effect size (ES) and p-value from Mann-Whitney Rank Sum test.

Supplementary Table Legends

Supplementary Table 1 – GEB across five individual breast cancer cohorts
Patient-level GEB in five recurrent amplicons for four individual cohorts: TCGA (primary invasive breast cancer), ICGC (primary invasive breast cancer), DCIS, Hartwig (metastatic breast cancer) and METABRIC (primary invasive breast cancer). Table includes subtype annotations, GEB in each of the four amplicons, genetic principal components and number of somatic SNVs for each cohort. METABRIC table additionally includes overall survival at five years and percent genome altered (PGA).

Supplementary Table 2 – Immune landscape of high *vs* low GEB tumors
Immune transcriptomic features from Thorsson *et al.* (*42*) for HER2+ and ER+ high *vs* low GEB tumors.