


SOFTWARE

Open Access



# Analysis of KIR gene variants in The Cancer Genome Atlas and UK Biobank using KIRCLE

Galen F. Gao<sup>1,2</sup>, Dajiang Liu<sup>3</sup>, Xiaowei Zhan<sup>4\*</sup> and Bo Li<sup>2\*</sup> 

## Abstract

**Background:** Natural killer (NK) cells represent a critical component of the innate immune system's response against cancer and viral infections, among other diseases. To distinguish healthy host cells from infected or tumor cells, killer immunoglobulin receptors (KIR) on NK cells bind and recognize Human Leukocyte Antigen (HLA) complexes on their target cells. However, NK cells exhibit great diversity in their mechanism of activation, and the outcomes of their activation are not yet understood fully. Just like the HLAs they bind, KIR receptors exhibit high allelic diversity in the human population. Here we provide a method to identify KIR allele variants from whole exome sequencing data and uncover novel associations between these variants and various molecular and clinical correlates.

**Results:** In order to better understand KIRs, we have developed KIRCLE, a novel method for genotyping individual KIR genes from whole exome sequencing data, and used it to analyze approximately sixty-thousand patient samples in The Cancer Genome Atlas (TCGA) and UK Biobank. We were able to assess population frequencies for different KIR alleles and demonstrate that, similar to HLA alleles, individuals' KIR alleles correlate strongly with their ethnicities. In addition, we observed associations between different KIR alleles and HLA alleles, including HLA-B\*53 with KIR3DL2\*013 (Fisher's exact FDR =  $7.64e-51$ ). Finally, we showcased statistically significant associations between KIR alleles and various clinical correlates, including peptic ulcer disease (Fisher's exact FDR = 0.0429) and age of onset of atopy (Mann-Whitney  $U$  FDR = 0.0751).

**Conclusions:** We show that KIRCLE is able to infer KIR variants accurately and consistently, and we demonstrate its utility using data from approximately sixty-thousand individuals from TCGA and UK Biobank to discover novel molecular and clinical correlations with KIR germline variants. Peptic ulcer disease and atopy are just two diseases in which NK cells may play a role beyond their "classical" realm of anti-tumor and anti-viral responses. This tool may be used both as a benchmark for future KIR-variant-inference algorithms, and to better understand the immunogenomics of and disease processes involving KIRs.

**Keywords:** Natural killer cells, Killer immunoglobulin receptors, Immunogenomics

## Background

Natural killer (NK) cells are an important component of the innate immune system that classically play an important role in the body's anti-tumor and anti-viral responses. In addition to their functions in these processes, recent research has further implicated their involvement in a much wider range of pathological processes that include cardiac, metabolic, oral, and gastrointestinal diseases [1–4]. While they represent only a small minority of circulating lymphocytes (10–15%), NK cells

\*Correspondence: xiaowei.zhan@utsouthwestern.edu; bo.li@utsouthwestern.edu

<sup>2</sup> Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA

<sup>4</sup> Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA

Full list of author information is available at the end of the article



nonetheless are considered to be the immune cell subtype most effective at monitoring and clearing diseased cells from the body [5].

As one mechanism to distinguish healthy host cells from infected or tumor cells, NK cells employ killer immunoglobulin receptor (KIR) proteins on their membrane surfaces to bind to and recognize Human Leukocyte Antigen Class I (HLA-I) complexes on the surface of their target cells. Fifteen KIR genes and 2 KIR pseudogenes have been discovered [6]. These 15 genes may broadly be categorized into either activating KIRs, which promote NK cell activation and induce killing of the target cell on receptor stimulation, or inhibitory KIRs, which prevent NK cell activation and spare the target cell upon ligand binding. Inhibitory KIRs generally possess long cytoplasmic tails and are denoted with an L (KIR2DL1, KIR2DL2, KIR2DL3, KIR2DL5A/B, KIR3DL1, KIR3DL2, and KIR3DL3), whereas activating KIRs generally possess short cytoplasmic tails and are denoted with an S (KIR2DS1, KIR2DS2, KIR2DS3, KIR2DS4, KIR2DS5, and KIR3DS1); however, KIR2DL4 uniquely among the 15 KIR genes possesses both activating and inhibitory functions [7]. Modulation of NK cell activity, and thus susceptibility or resistance to various pathologies, likely depends strongly on the binding properties and interactions between KIR and HLA-I molecules. A high level of diversity in NK cell activity and its outcomes may be achieved largely through four different mechanisms: KIR recognition of highly distinct subsets of HLA-I allotypes, combination of KIRs into distinct haplotypes in different individuals, stochasticity of KIR expression on the surface of individual NK cells, and allelic polymorphism of individual KIR genes [8]. In this manuscript, we primarily explore the last of these mechanisms and its downstream effects on disease susceptibility by performing KIR allele inference using next-generation sequencing (NGS) data.

Previous attempts to perform KIR genotyping at the individual gene level have either (1) relied on specially prepared primers and amplicon design, (2) required manual review as part of the algorithm, (3) utilized an experimental platform completely different from NGS, or (4) have merely assessed KIR gene presence or deletion rather than detected single-nucleotide-variants [9–12]. Given the rise and modern prevalence of NGS, especially with the recent releases of Whole Exome Sequencing (WES) data for large datasets including UK Biobank and The Cancer Genome Atlas (TCGA), there is a strong need for a fully automated pipeline that can detect single-nucleotide variants of these KIR genes using aligned WES data. In this work, we have developed and characterized the performance of a fully automated algorithm for accurate inference of KIR gene alleles from WES data: “KIR CaLLing by Exomes” (KIRCLE). To demonstrate the

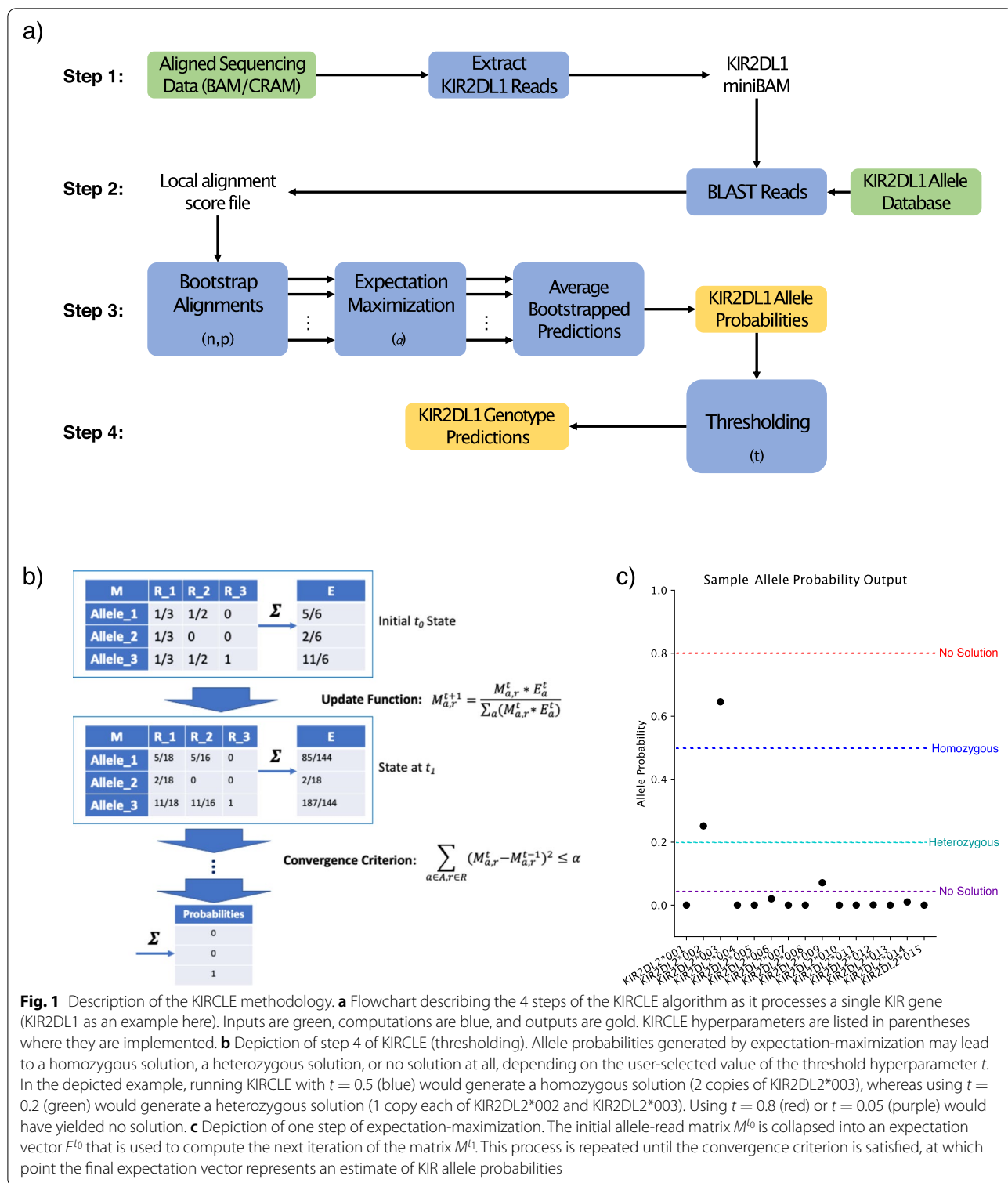
utility of such an automated KIR genotyper, after running KIRCLE on 10,332 TCGA and 49,953 UK biobank exome samples, we discovered several novel correlations between KIR allele calls and other molecular and clinical features in these two datasets. Our work represents the first large-scale genetic analysis to elucidate pathologic and immunologic associations with human natural killer cells and provides an unprecedented resource for future investigations into the functionality of different KIR alleles.

## Results

### KIRCLE workflow description

KIRCLE is an allele inference algorithm that uses aligned WES data in the form of a BAM or CRAM file to generate probability estimates for each KIR allele, as well as genotype predictions for each KIR gene. KIRCLE consists of 4 major steps: pre-processing, local alignment with BLAST, bootstrapped expectation-maximization, and thresholding (Fig. 1a).

- 1) In pre-preprocessing, KIRCLE first extracts all WES reads that map to the genomic coordinates of the KIR genes on chromosome 19q13.4 and writes these reads to fifteen separate files—one for each KIR gene.
- 2) Next, KIRCLE uses nucleotide BLAST to perform local alignment on each KIR gene’s collection of reads against a database of variants belonging to that particular KIR gene. In the IPD-KIR Database v2.8.0, 908 different alleles spanning the 15 KIR genes are documented, of which 535 represent distinct coding variants. KIRCLE then filters out alignments with less than 100% identity matches to documented KIR alleles.
- 3) KIRCLE then bootstraps the BLAST-identified alignments with 100% identity matches to KIR alleles and uses an expectation-maximization (EM) algorithm, with convergence hyperparameter  $\alpha$ , to generate allele probability estimates from these collections of alignments (Fig. 1c).  $n$  bootstraps of fraction  $p$  of all 100%-identity alignments are computed in this manner. The bootstrapped allele probability estimates are then averaged together to determine a final probability estimate for each allele. This bootstrapping is helpful in countering the EM algorithm’s tendency to converge to local minima representing homozygous solutions based on small differences in initial alignment data.
- 4) Finally, KIRCLE uses a thresholding algorithm to convert each KIR gene’s set of allele probability estimates into homozygous or heterozygous genotype calls, depending on the number of alleles that exceeded a heuristically determined threshold  $t$



(Fig. 1b). Depending on the user's selection of hyperparameter  $t$  and the EM algorithm's outputted allele probabilities, the resulting genotype solution may be either homozygous, heterozygous, or non-existent.

Final workflow outputs from KIRCLE include a table of allele probability estimates, a table of genotype calls, and a list of runtime hyperparameters.

### Hyperparameter determination and validation for KIRCLE

KIRCLE requires the use of 4 hyperparameters:  $\alpha$  (the convergence threshold for expectation-maximization),  $n$  (the number of bootstraps to perform),  $p$  (the proportion of reads to use in each bootstrap), and  $t$  (the threshold used to convert KIR allele probabilities to binary KIR genotype calls). Of these, choices regarding  $p$  and  $t$  represent the greatest and most direct potential sources of variability in KIRCLE's accuracy. Using one randomly selected sample from UK Biobank, we were able to characterize KIRCLE's performance, as measured by the Shannon entropy of the inferred genotypes, across different values of  $p$  (from 0.2 to 0.8) and  $t$  (from 0.05 to 0.40). At each set of hyperparameters tested, we performed 500 iterations of KIRCLE on one arbitrarily selected sample in UK Biobank, collected the 500 genotype outputs, and empirically computed the log<sub>2</sub> Shannon entropy of the genotype solutions for each KIR gene. An ideal genotype caller would be consistent and call the same solution for the same input, resulting in a "genotype-entropy" of 0. For many KIR genes, such as KIR2DL1 and KIR2DL4, contour maps of the resulting entropies revealed that KIRCLE was largely self-consistent, with little variability of output (genotype-entropy of 0) across a wide spectrum of hyperparameter values (Fig. 2a, b). This pattern was recapitulated in the majority of KIR genes, suggesting respectable consistency of KIRCLE output across multiple KIR genes (Supplementary Fig. S1a–j). For all subsequent analyses in this manuscript, hyperparameter values of  $\alpha=1e-5$ ,  $n=100$ ,  $p=0.5$ , and  $t=0.25$  were used.

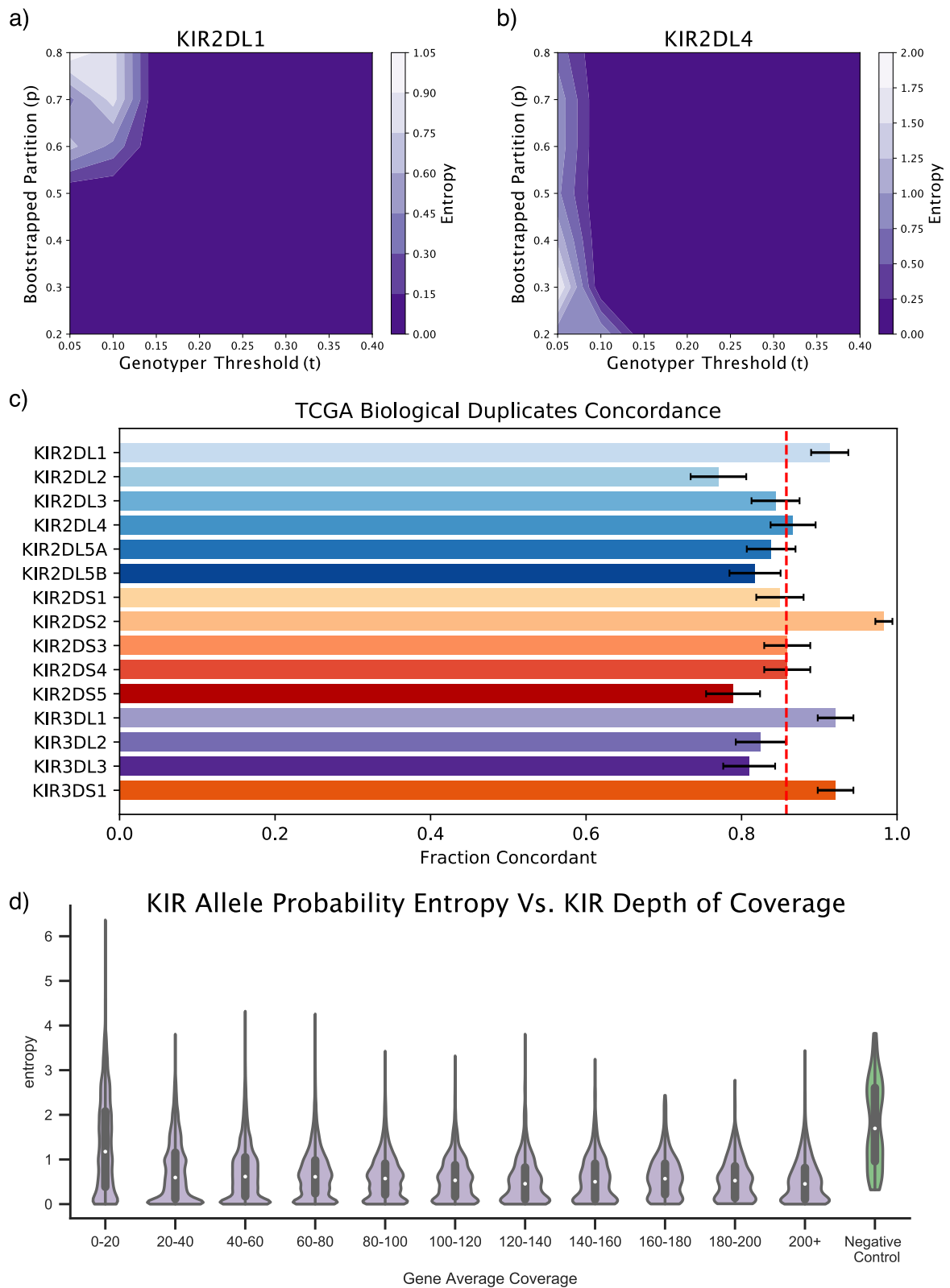
Next, to establish the accuracy of our algorithm, we assessed the concordance of KIRCLE-generated genotypes between TCGA biological replicates. Of 10,332 exomes in TCGA, 1,062 were present twice as biological replicates and thus used in this analysis. We determined that 85.8% of genotype solutions called by KIRCLE across all KIR genes were concordant between replicates (Supplementary Fig. S2a). We defined solutions to be concordant if the genotype inferred by KIRCLE in one sample was identical to that inferred in its replicate. Genes with the highest concordance between replicates were KIR2DS2 (98.3%), KIR3DL1 (92.1%), and KIR3DS1 (92.1%), whereas genes with the lowest concordance between replicates were KIR2DL2 (77.0%), KIR2DS5 (78.9%), and KIR3DL3 (81.0%) (Fig. 2c). Given the observed difference between KIRs with the highest

concordance (e.g., KIR2DS2) and those with the lowest concordance (e.g., KIR2DL2), we sought to explain these differences in accuracy by analyzing the degree of sequence similarity between different alleles of each KIR. We assessed sequence similarity by performing a multiple sequence alignment (MSA) on all alleles of a KIR using Clustal Omega [13] and then measuring the mean phylogram distance between all coding variants (Supplementary Fig. S2b–c). We noted a strong positive correlation (Spearman's  $\rho=0.631$ ) between the mean distance between alleles of a given KIR and the observed concordance between that KIR's genotype calls among TCGA biological replicates, indicating that the higher levels of sequence similarity among alleles of KIRs such as KIR2DL2, KIR2DS5, and KIR3DL3 are likely to account for their lower rates of observed concordance.

Finally, we investigated whether KIRCLE is robust against differences in depth of sequencing. To do so, we compared the ambiguity of KIRCLE's output, quantified as the Shannon entropy of generated KIR allele probabilities, across TCGA samples with different depths of sequencing. Low ambiguity in KIR allele calling results in KIR allele probabilities of either 1 for a single allele and 0 for all other alleles (reflecting a homozygous genotype) or 0.5 for two different alleles and 0 for all other alleles (reflecting a heterozygous genotype), leading to entropies of either 0 or 1 respectively. Conversely, high ambiguity in KIR allele calling will lead to a more uniform distribution of KIR allele probabilities, leading to entropies higher than 1. For each TCGA sample, we measured both the average coverage and the KIR-allele-probability entropies for each KIR gene. Binning samples by their average coverages, we observed that allele probability entropies—and thus the ambiguity of KIR allele probabilities—are notably increased only at very low coverages (<20 average depth of coverage at the KIR gene locus). Furthermore, as negative controls, 20 "pseudo-BAMs" were generated by randomly sampling reads mapping to KIR gene loci from 50 randomly selected BAMs in TCGA. Pseudo-BAMs were generated with an average read depth commensurate with their constituent BAMs. After applying KIRCLE to these pseudo-BAMs, their resulting allele probability entropies were much higher (median=1.70; IQR 0.958–2.61) than a significant majority of actually observed entropies for all TCGA samples, regardless of the depth of coverage (Fig. 2d). Moreover,

(See figure on next page.)

**Fig. 2** KIRCLE accuracy and consistency validation. **a** Contour plot demonstrating the effect of varying the bootstrap-proportion ( $p$ ) and threshold ( $t$ ) hyperparameters on KIR2DL1 allele inference, as measured by empirical calculation of the inferred genotypes' entropy. **b** KIRCLE's performance on KIR2DL4 allele inference was similarly characterized. **c** Fraction of each KIR gene's KIRCLE-inferred allele genotypes that were called identically between 531 samples and their biological replicates in TCGA. Error bars represent the normal approximation confidence intervals. **d** TCGA sample coverages (binned) versus TCGA sample allele probability entropies for all 15 KIR genes. The allele probability entropies of a set of 20 "pseudo-BAMs" (green) are presented as negative controls



**Fig. 2** (See legend on previous page.)



despite differences in average depth of coverage at different KIR gene loci, average KIR allele entropies between different KIR genes largely remained constant (Supplementary Fig. S2b). Overall, KIRCLE demonstrated a high level of consistency while being able to call a diverse set of KIR genotype solutions and is robust to the effects of low depth of coverage.

#### KIR allele comparisons between TCGA and UK Biobank

After benchmarking KIRCLE using internal quality control metrics, we assessed KIRCLE's performance by comparing its allele predictions in TCGA to its allele predictions in UK Biobank. We first compared the frequencies of different KIR alleles in TCGA with their frequencies in UK Biobank among Caucasian individuals in both datasets, in order to mitigate race as a confounding variable affecting observed allele frequency. For each KIR gene, we ranked its alleles by frequency in both TCGA and UK Biobank and then computed the Spearman correlation coefficient between the allele frequencies in the two datasets (Fig. 3a). We noted that all KIR genes displayed positive correlation coefficients and that the vast majority of KIR genes demonstrated highly similar distributions of allele frequencies between TCGA and UK Biobank (Spearman's  $\rho=0.802$ ). Direct comparison of all KIR alleles ranked by frequency also demonstrated high consistency between the two cohorts' Caucasian subpopulations (Fig. 3b). Both UK Biobank and TCGA are largely composed of Caucasians (81.4% and 93.3% of the individuals analyzed in TCGA and UK Biobank respectively).

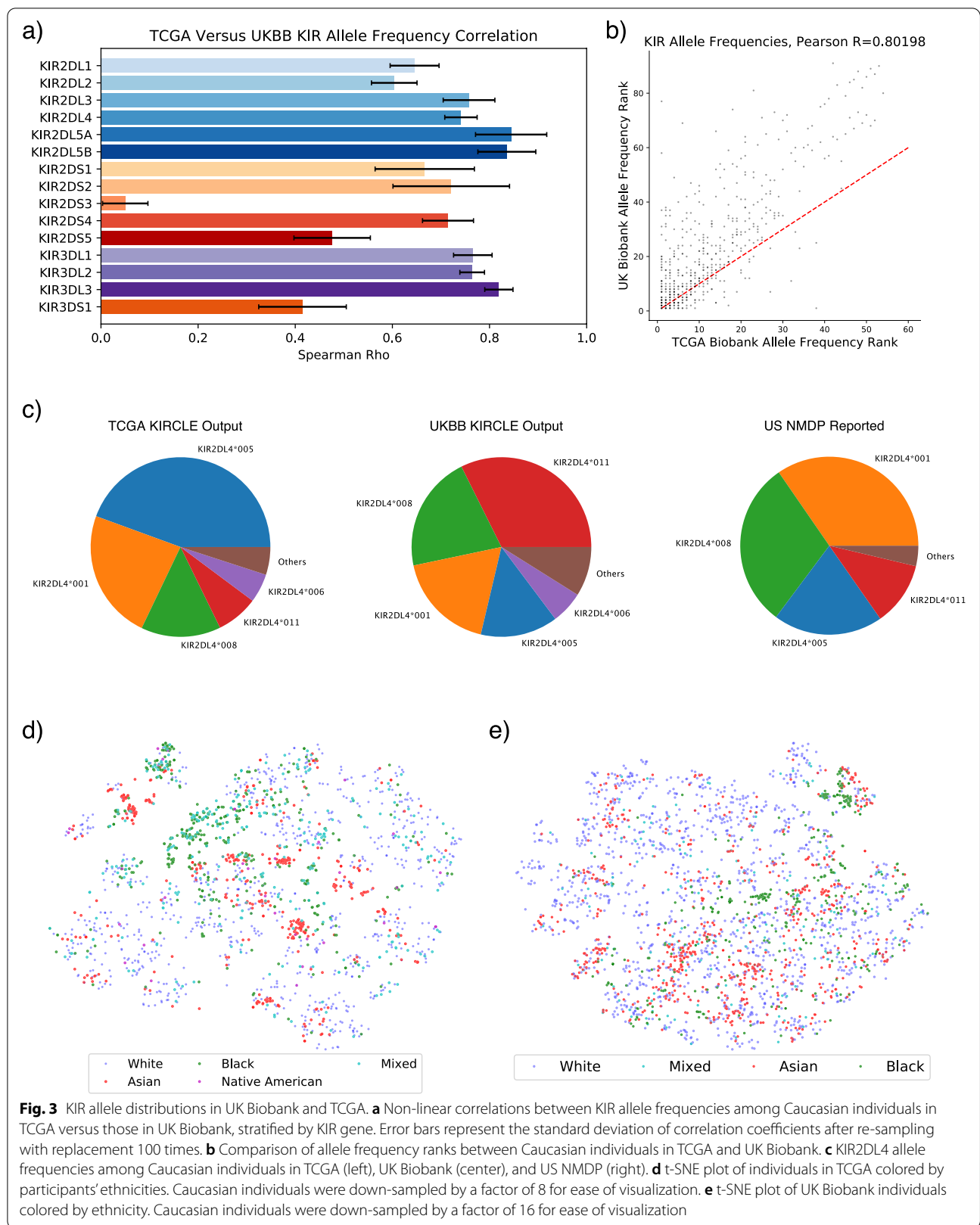
Additionally, we were able to further validate observed KIR allele frequencies for certain KIR genes using allele frequency data from the US National Marrow Donor Program (NMDP), as reported by the Allele Frequency Net Database [14]. We used the NMDP dataset because the subjects in this cohort are predominantly Caucasians, similar to the TCGA patients. For KIR2DL4, the four most frequent KIR2DL4 alleles reported by the NMDP were KIR2DL4\*001, KIR2DL4\*008, KIR2DL4\*005, and KIR2DL4\*011 (34.6%, 30.2%, 19.9%, and 11.6% respectively). We were able to recapitulate these four alleles as the most frequent KIR2DL4 alleles in both TCGA and UK Biobank's Caucasian subpopulations, albeit in a different order for each dataset (Fig. 3c). In TCGA, KIR2DL4\*005 was the most frequent allele, followed by KIR2DL4\*001, KIR2DL4\*008, and KIR2DL4\*011 (44.4%, 23.5%, 14.4%, and 7.55%). In UK Biobank, this order was reversed with KIR2DL4\*011 being the most frequent allele, followed by KIR2DL4\*008, KIR2DL4\*001, and finally KIR2DL4\*005 (32.4%, 21.0%, 18.0%, and 13.9%). Further validation of allele frequencies against the NMDP was also performed for the alleles of KIR3DL2. The most frequent KIR3DL2

allele in a population of 75 Caucasians was KIR3DL2\*002 (26.1%), followed by KIR3DL2\*001 and KIR3DL2\*007 (21.0% and 18.8%) [15]. While KIR3DL2\*002 was found at similarly high frequencies in both TCGA (9.37%) and UK Biobank (9.27%) as the 4th and 3rd most frequent KIR3DL2 alleles respectively, KIR3DL2\*001 and KIR3DL2\*007 were much lower ranked at 9th and 13th in TCGA and 8th and 1st in UK Biobank respectively. However, these are still fairly well-represented alleles at 4.95% and 2.78% frequency in TCGA and 5.00% and 13.3% frequency in UK biobank respectively. Furthermore, considered overall, KIR3DL2 allele frequency ranks in TCGA and UK Biobank still demonstrate positive correlations with the allele frequency ranks observed in the NMDP (Supplementary Fig. S2c). Despite slight numerical differences, confirmation of the status of the most frequent alleles in these two KIR genes increases our confidence in KIRCLE's ability to infer KIR alleles from WES data accurately.

In addition to validating population frequencies of KIR alleles, we also examined patterns of KIR allele co-expression and dependence. As KIRCLE assesses for the presence of 535 KIR alleles over 15 KIR genes, the KIR genotype of each sample in TCGA and UK Biobank may be represented as a point in 535-dimensional "KIR-space." We first used t-distributed stochastic neighbor embedding (t-SNE) to perform dimensionality reduction and thus visualize the distribution of individuals in TCGA in 2 dimensions [16]. When we colored this t-SNE map using individuals' SNP-inferred ethnicities [17], we observed that different ethnicities cluster together and are non-uniformly distributed (Fig. 3d). In particular, African Americans and—to a lesser extent—Asian Americans in TCGA formed clusters that were often very distinct from the Caucasian majority. Similar analyses performed in UK Biobank recapitulated this non-random distribution of KIR genotypes and confirmed the non-uniform distribution and clustering of those who self-identified their ethnicity as "Black" or "Asian" (Fig. 3e). Of particular note, the "Asian" population in TCGA comprises those of East Asian descent, whereas the "Asian" population in UK Biobank largely comprises those of South Asian descent (with major subcategories of Indians, Pakistanis, and Bangladeshis). However, both groups of Asians clustered distinctly and separately from the Caucasian majority to some extent in both datasets.

#### KIR allele associations with HLA alleles

As it is known that HLA and KIR bind to each other in an allele-specific way, we posited that strong correlations may also exist between KIR alleles and HLA alleles on the population level, due to a known co-evolution event in humans [18]. Using HLA types imputed by the



HLA\*IMP:02 algorithm and subsequently released by UK Biobank [19], we observed 326 significantly associated pairs of KIR alleles with HLA alleles in the UK Biobank data (Fig. 4a, b, Supplementary Fig. S3a). Many of these associations belonged to a set of particularly common HLA alleles (e.g., HLA-B\*53) or KIR alleles (e.g., KIR3DL3\*005). Furthermore, we also note that the majority (73.0%) of significant associations are positive. We speculate that these associations reflect changes in direct physical interactions between HLA and KIR alleles, which result in co-selection due to an advantageous increase in fitness for individuals with these combinations of KIR and HLA alleles. Particularly visually striking examples of positive and negative associations between KIRs and HLAs include KIR3DL3\*005 with HLA-A\*74 (Fisher's exact FDR=7.43e-43; odds ratio=55.1) and KIR2DL3\*002 with HLA-A\*36 (Fisher's exact FDR=1.71e-12; odds ratio=0.0727), respectively (Fig. 4c). Additionally, when examining the t-SNE coordinates of individuals with HLA alleles such as HLA-B\*42, we observed a non-uniform distribution and clustering of these samples that closely mirrors the distribution of samples when labeled by ethnicity (Fig. 4d).

While these findings may support the biological link between these two classes of molecules and shed additional light onto which particular HLA alleles may have evolved in parallel with particular KIR alleles, they also raise the possibility that our observed associations are driven by population stratification according to ethnicity. In order to disentangle the effects of this stratification on associations between HLA and KIR alleles, we re-attempted the analysis using only Caucasian individuals in UK Biobank, while testing only KIR alleles with >1% allele frequency in UK Biobank (Fig. 4e). While this analysis unveiled a much smaller subset of HLA-KIR associations, we noted 3 significant associations: HLA-C\*17 with KIR2DS4\*016 and HLA-B\*41 with KIR2DL4\*011 and KIR2DS4\*016. Notably, both KIR2DS4 and KIR2DL4 have NK-cell-activating activity, and all three are affiliated with a negative odds ratio. These results indicate that HLA-C\*17 and B\*41 could be true activation ligands for KIR2DS4 and KIR2DL4, and their interactions may induce NK responses that impose negative selection pressure on individuals bearing both alleles.

Although TCGA is a much smaller dataset than UK Biobank, we were able to use TCGA to discover a smaller set of correlations between HLA alleles and KIR alleles after filtering out KIR alleles with <1% allele frequency in TCGA to improve our Bonferroni correction factor (Supplementary Fig. S3b). HLA allele calls for samples in TCGA were made using POLYSOLVER [20]. In particular, KIR2DL2\*003, KIR3DL2\*013, and KIR3DL3\*008 were strongly positively associated with HLA-B\*46, HLA-B\*53,

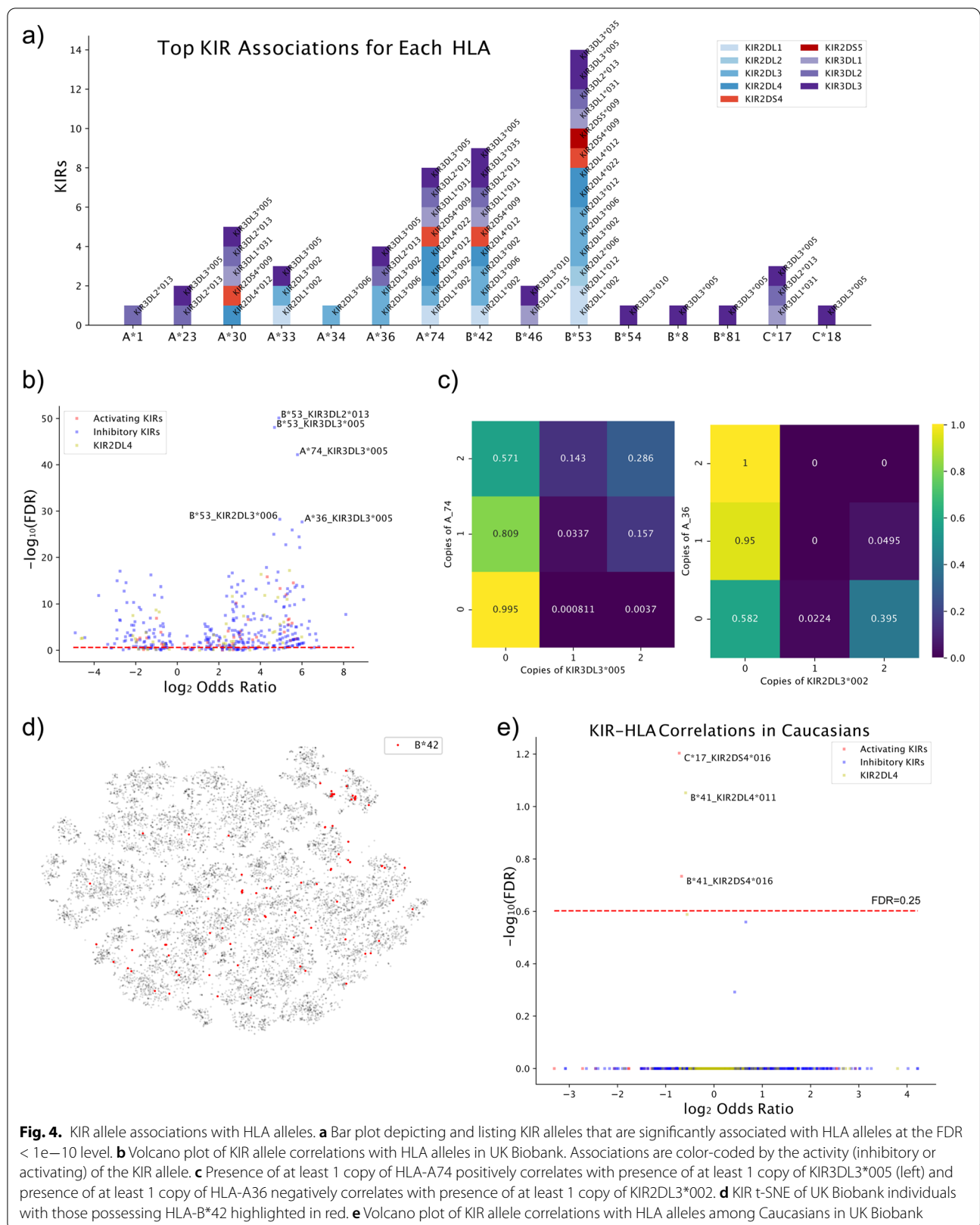
and HLA-C\*15 respectively at the FDR < 0.25 level. The HLA-B\*53 association with KIR3DL2\*013, notably, was the most significant HLA-KIR association discovered in UK Biobank. However, when we re-attempted the analysis using only Caucasian individuals in TCGA to eliminate population stratification by ethnicity as a potential confounding factor, all significant associations between KIR and HLA alleles disappeared after Bonferroni correction. In summary, after correction of population stratifications, we found few significant associations between activating the KIR gene and HLA alleles. The absence of significant associations between inhibitory KIR genes and HLA alleles might suggest weaker selective pressure for KIR alleles, possibly due to the multiple redundant mechanisms inhibiting NK cell activation [21].

#### KIR allele associations with clinical correlates

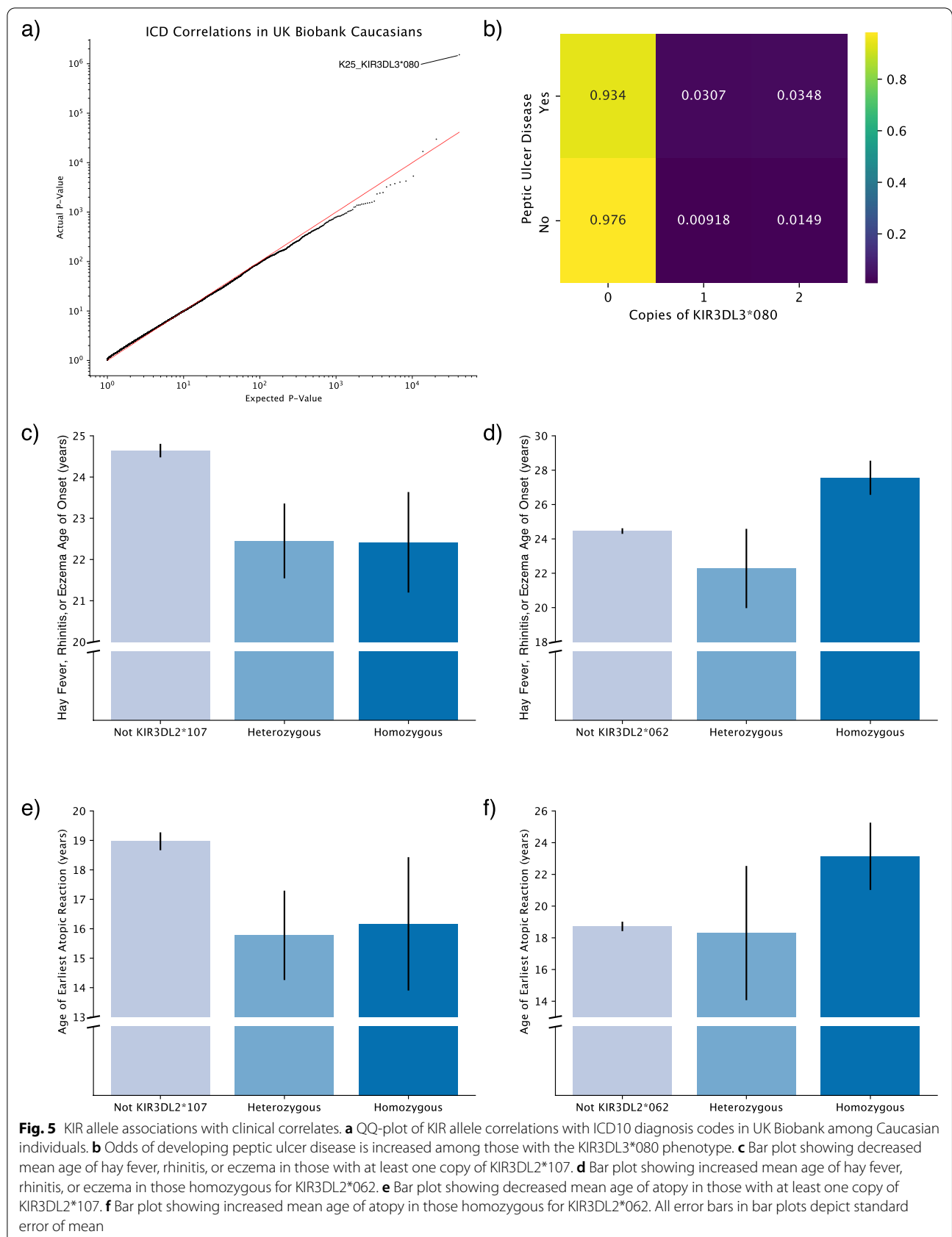
In addition to correlations with HLA alleles, we searched for KIR allele correlations with clinical features. We first examined KIR allele correlations with individuals' medical diagnoses documented in UK Biobank, as encoded by the 10th revision of the International Statistical Classification of Diseases (ICD10). To minimize the number of under-powered tests we performed, we attempted correlations only with KIR alleles represented at over 1% frequency in UK Biobank. Additionally, we excluded all diseases primarily associated with external causes, including accidents, injuries, and nutritional deficiencies, as well as obstetric and psychiatric diseases among others. Of note, this list of exclusions includes infectious diseases, which despite having a strong biological basis for association with KIR alleles, require exposure to a pathogen, which is largely driven by individuals' environmental circumstances. Strikingly, the only associations that remained significant at the FDR < 0.25 level were those associated with sickle-cell anemia (ICD10 D57) or with uterine leiomyomas (ICD10 D25), both diseases that disproportionately affect black people [22]. However, positing a direct biological mechanism behind these associations likely would represent a third-cause fallacy, as blacks are statistically more likely to possess both KIR alleles enriched in black populations as well as either the sickle-cell trait or uterine leiomyomas.

Thus, we next narrowed our analysis to investigate only those individuals who self-identified as Caucasian. While the vast majority of correlations failed false-discovery-rate correction, we discovered a significant correlation between the KIR3DL3\*080 allele and ICD10 K25—peptic ulcer disease (PUD) (Fisher's exact test, FDR= 0.0429; Fig. 5a). Whereas those without KIR3DL3\*080 had merely a 1.04% chance of being diagnosed with PUD, patients with KIR3DL3\*080 had a 2.90% chance of being diagnosed with PUD, representing a 2.8-fold increase in





**Fig. 4.** KIR allele associations with HLA alleles. **a** Bar plot depicting and listing KIR alleles that are significantly associated with HLA alleles at the FDR < 1e−10 level. **b** Volcano plot of KIR allele correlations with HLA alleles in UK Biobank. Associations are color-coded by the activity (inhibitory or activating) of the KIR allele. **c** Presence of at least 1 copy of HLA-A74 positively correlates with presence of at least 1 copy of KIR3DL3\*005 (left) and presence of at least 1 copy of HLA-A36 negatively correlates with presence of at least 1 copy of KIR2DL3\*002. **d** KIR t-SNE of UK Biobank individuals with those possessing HLA-B\*42 highlighted in red. **e** Volcano plot of KIR allele correlations with HLA alleles among Caucasians in UK Biobank



likelihood (Fig. 5b). No significant association was found between KIR3DL3\*080 and usage of ibuprofen, which would predispose individuals toward developing PUD (data not shown). Thus, if KIR3DL3\*080 predisposes an individual toward PUD, it likely does so through an alternative mechanism. Significant correlations with ICD10 diagnosis codes in Black and Asian populations were not observed, likely owing to the lower statistical power these smaller populations had.

Additionally, we explored correlations between KIR alleles with population frequency >1% and other clinical correlates besides ICD10 codes. When examining correlations with age of onset of several chronic diseases and conditions, we discovered that KIR3DL2\*107 was highly correlated with early age of onset of hay fever, rhinitis, or eczema in Caucasian individuals. Whereas individuals without KIR3DL2\*107 had an average age of onset of 24.7 years (IQR 12–35 years), those with at least one copy of KIR3DL2\*107 had an average age of onset of 22.4 years (IQR 11–30 years; two-sided Mann-Whitney FDR=0.0751; Fig. 5c). Moreover, an alternative allele of KIR3DL2, KIR3DL2\*062, was weakly associated with an increase in age of onset of hay fever, rhinitis, or eczema from 24.5 years (IQR 12–35 years) to 27.0 years (IQR 14–40 years; Mann-Whitney FDR=0.244; Fig. 5d). Later onset of these conditions was particularly pronounced in individuals with two copies of KIR3DL2\*062, with an average age of onset of 27.6 years (IQR 14–40 years). Together, these results suggest that polymorphisms in KIR3DL2 may play a key role in determining the age of onset of hay fever, rhinitis, and/or eczema.

Moreover, hay fever and eczema, in conjunction with allergic asthma, more broadly represent manifestations of atopy, the genetic predilection to trigger IgE-mediated (type I) hypersensitivity reactions following allergen exposure with increased  $T_H2$ -driven responses [23]. Thus, we next attempted to generalize this association to encompass atopy more broadly by examining KIR3DL2\*107 and KIR3DL2\*062's associations with age of onset of either asthma or hay fever, rhinitis, or eczema, using the age of onset of whichever condition occurred earliest in life for each individual. We observed the same association: individuals with at least one copy of KIR3DL2\*107 had an average age of onset of 15.9 years (IQR 6–20.75 years), whereas those without any copies of KIR3DL2\*107 had an average age of onset of 19.0 years (IQR 8–28 years; Mann-Whitney  $p=0.012$ ; Fig. 5e). Simultaneously, individuals with at least one copy of KIR3DL2\*062 (22.5 years; IQR 10.25–31.5 years), and particularly those with two copies of KIR3DL2\*062 (23.1 years; IQR 10.75–32.75 years), had later onsets of atopic reactions than those without KIR3DL2\*062 (18.7 years; IQR 7–27 years;

Mann-Whitney  $p=0.019$ ; Fig. 5f). Together, these findings suggest a potential biological mechanism either delaying or hastening onset of atopic reactions like hay fever, eczema, or asthma that involves KIR3DL2, and the KIR3DL2\*107 and KIR3DL2\*062 alleles in particular. In addition to atopic reactions, we also observed significant associations of KIR alleles with other clinical correlates, including dental and oral health, quantitative blood analysis, and waist circumference, suggesting potentially broad impact of natural killer functions in affecting diverse human traits (Supplementary Fig. 4).

Finally, to further explore the effects of KIR3DL2 polymorphism on age of atopy onset, we posited that each of the two aforementioned KIR3DL2 alleles follows either a dominant, semi-dominant, or recessive model of expression and then sought to determine which of these three models best explains the effect of KIR3DL2 genotype on age of atopy onset. In the recessive model, only a genotype homozygous for the KIR3DL2 allele in question contributes to a change in age of atopy onset from baseline. In contrast, in the dominant model, genotypes either homozygous or heterozygous for the KIR3DL2 allele in question contribute to changes in baseline age of atopy onset. Finally, in the semi-dominant model, homozygotes for the KIR3DL2 allele in question are twice as potent as corresponding heterozygotes in changing age of atopy onset from baseline. When assessed against each other using the UK Biobank data, the dominant model outperformed semi-dominant and recessive models of expression for KIR3DL2\*107, as measured by the Bayesian information criterion (−10.8145 versus −10.8151 and −10.8172, respectively). Meanwhile, expression patterns of KIR3DL2\*062 instead favored the recessive model over the semi-dominant and dominant models of expression for KIR3DL2\*062, as measured by the Bayesian information criterion (−10.8138 versus −10.8141 and −10.8146, respectively). In summary, our analysis indicated that KIR3DL2\*107 may “override” other alleles and thus present with a dominant phenotype, whereas KIR3DL2\*062 may be weaker than other KIR3DL2 alleles and thus present with a recessive phenotype.

## Discussion

The fifteen KIR genes represent a polymorphic set of immune modulators with an array of potential effects on immune and clinical phenotypes. In this manuscript, we have developed, characterized, and implemented our algorithm KIRCLE, uncovering multiple correlations between KIR alleles and other features in TCGA and UK Biobank.

### KIR alleles associated with HLA alleles

Class I HLAs represent well-known binding partners of KIRs. Thus, any change in either the KIR binding site or the HLA binding site that alters their affinities to each other may be expected to modulate NK cell activation or inhibition. However, as both HLA and KIR loci are highly polymorphic, it has historically been challenging to determine their matches through low-throughput experimental approaches or through small-scale computational analyses. By using KIRCLE and a large cohort of UK Biobank data, we were able to observe several statistically significant associations between KIR and HLA alleles. These observed KIR-HLA associations may be indicative of selective pressures for these receptors to co-evolve to maintain appropriate levels of NK cell activity. Specifically, we observed that Caucasian individuals in UK Biobank with KIR2DS4\*016 and KIR2DL4\*011, two KIR alleles with activating activity, have a lower frequency of HLA alleles HLA-C\*17 and HLA-B\*41 than individuals without these alleles. We may posit that this “anti-correlation” possibly represents evidence of a potential intolerance of lethal NK cell hyperactivity, leading to the observed underrepresentation of individuals with these particular KIR-HLA allele combinations. However, such an explanation is purely speculative, with further mechanistic studies required to validate the observed frequencies and either to confirm or to refute this hypothesis.

Several factors affect the results of our analyses. In addition to observing differences in KIR allele frequencies among those with different HLA alleles, we also observed differences in KIR allele frequencies among different ethnic populations, which are already known to possess different HLA allele frequencies [24]. While this combination of observations may reflect common selective pressures historically experienced by these ethnic populations which then may have forced KIR and HLA alleles to co-evolve, it likely also points to ethnicity as a confounding factor in this purely correlative study. We attempted to account for this potential confounder by repeating our analysis on only Caucasian individuals within UK Biobank. Such an analysis yielded a much smaller set of statistically significant KIR-HLA associations. Finally, it is possible that the KIR allele cohorts as inferred by KIRCLE are not well-defined enough to fully elucidate the underlying KIR-HLA interactions driving any observed associations. Future work would be expected to resolve these cohorts more accurately and thus better detect KIR-HLA interactions as well as their effects on immune and cancer-related outcomes.

### KIR allele associations with peptic ulcer disease and atopic reactions

Previous genome-wide association studies of PUD have largely been performed in East Asian populations and did not uncover any associations between KIR polymorphism and either PUD or *H. pylori* infection [25, 26]. However, NK cells are known to be present in the gastric and duodenal mucosa and have been shown to be directly activated by *H. pylori* bacteria to produce IFN- $\gamma$  and trigger an immune response [27]. Our result builds upon these existing known interactions and hypothesizes that KIR3DL3\*080 may increase susceptibility to PUD through modulating NK cells' natural response to *H. pylori*.

Furthermore, we uncovered evidence for a potential association between age of presentation of atopy and two different variants of KIR3DL2: KIR3DL2\*107 and KIR3DL2\*062. At least one copy of one of these two variants is present in 7.89% of the Caucasian population in UK Biobank. Indeed evidence exists for NK cells' involvement in atopic and autoimmune diseases of the skin (i.e. eczema), even if the details of this involvement remain unclear [28], and increasing support has been seen for their role in allergen-specific immune suppression, Th1 cell generation, and Ig production [29]. Our finding that KIR3DL2\*107 is associated with earlier presentation of atopy and that presence of KIR3DL2\*062 associates with delayed presentation potentially further points to a role specifically for KIR3DL2 in regulating NK cell activity as it contributes to these diseases. One possible explanation for the opposite directions of impact on age of onset is that KIR3DL2\*107 is stronger than other alleles and thus presents with a dominant phenotype, whereas KIR3DL2\*062 is weaker than others and thus results in a recessive phenotype. As preliminary evidence supporting this explanation, we demonstrated that a dominant model of expression best fits KIR3DL2\*107, while a recessive model of expression best fits KIR3DL2\*062.

However, it is worth emphasizing that even though our observed clinical correlations remained significant after correction for multiple hypothesis testing, it remains unclear to what extent the identified KIR alleles are the direct cause of these phenotypic changes. Moreover, NK cells *in vivo* utilize a combination of multiple KIRs and other receptors to interact with their target cells, whereas our current analyses examine bulk correlations with KIR genotypes and lack the single-cell resolution required to investigate the combinatorial complexity of actual KIR expression on the surface of NK cells *in vivo*. Thus, a more nuanced and more adequately powered study of the effects of multiple KIRs in combination at single-cell resolution may better resolve the biological interactions

of KIRs that ultimately functionally cause these downstream clinical outcomes.

### Limitations and future directions

While we were able to use biological replicates in TCGA to benchmark the accuracy of KIRCLE and compare our population-level estimates of KIR allele frequencies to prior estimates of KIR allele frequencies as reported by the US NMDP to validate our KIR genotype predictions, we were unable to carry out any experimental validation to benchmark its accuracy more directly. While our current analysis includes measures of the Shannon entropies of KIRCLE-predicted allele probabilities and concordance of predicted genotypes between biological replicates within TCGA as measures of the precision of our algorithm, such measures nevertheless remain susceptible to systematic biases in either the methodology or the nature of the datasets used (e.g., WES). In particular, our current analysis does not compare our KIR genotype predictions to gold-standard KIR genotypes as assessed by Sanger sequencing against each KIR gene's locus on chromosome 19 in live biological samples. Such comparison to gold-standard genotype predictions will be required to evaluate the overall accuracy of any KIR genotype inference algorithm more definitively.

Additionally, after generating KIR genotype predictions, our downstream correlations all represented univariate analyses, due to the relatively low abundance of individual KIR alleles. While such a simplistic analysis is suitable for a first-pass search for potential direct associations, more nuanced future analyses of clinical associations with KIR alleles will need to account for confounding factors beyond human genetics to determine individuals' susceptibility to diseases, including individuals' living or occupational environments, full medical histories, lifestyles, and much more. Such analyses may help uncover any functional roles KIRs play in these processes and will likely become available with sufficient statistical power when UK Biobank fulfills its mission to sequence all 500,000 individuals. Meanwhile, the studied sample sizes of ~50,000 individuals in UK Biobank and ~10,000 individuals in TCGA are relatively underpowered to discover associations between all combinations of 535 KIR alleles with thousands of molecular and clinical correlates. As such, we employed a relatively liberal threshold of  $FDR < 0.25$  in our downstream analyses. While an FDR threshold of 25% indicates that reported results are likely to be valid 3 out of 4 times and is appropriate in the context of exploratory, discovery-driven analyses, our reported findings in this manuscript nevertheless should be interpreted merely as interesting candidate hypotheses that require further validation in future work.

Furthermore, Caucasian individuals are heavily over-represented in both the TCGA and UK Biobank cohorts, and thus our downstream analyses have largely been suitably powered to investigate only those KIR alleles that are well represented among Caucasians. Future studies will be needed to use more racially diverse cohorts to analyze KIR alleles that are more frequently represented in other ethnicities.

Finally, as mentioned at the outset, NK cell activity is modulated by a number of factors outside of individual KIR genes polymorphism, including the subset of HLA-Is they recognize, the distinct combinations of genes that constitute the individual's KIR haplotype, and stochasticity in KIR expression on the surface of individual NK cells. Indeed, over 40 distinct KIR haplotypes, each composed of at least seven KIR genes, have been documented in the human population [30]. Variation of any of these additional factors may further affect NK cell function and ideally would be explored in conjunction with KIR polymorphism at the individual gene level in future studies.

In conclusion, our work has generated KIR allele predictions for TCGA and UK Biobank that will be invaluable for future studies of NK cells in these populations, uncovered multiple potential novel associations between KIR gene variants and clinical and molecular features, and has paved the way for future investigation into the role of KIRs in immunologic response and human disease. We hope that our algorithm can serve as a benchmark for future algorithms that will perform KIR genotyping and that others may use our algorithm to better understand the immunologic and pathologic processes surrounding KIR genes.

### Conclusions

We have developed KIRCLE, a first-of-its-kind fully automated computational pipeline for the inference of germline variants of the highly polymorphic killer-cell immunoglobulin-like receptor (KIR) genes from whole exome sequencing data. We demonstrate the utility of such an algorithm by using KIRCLE to infer germline KIR variants in approximately sixty-thousand individuals in The Cancer Genome Atlas and UK Biobank and then discover novel molecular and clinical correlations with these variants. This work represents the first large-scale genetic analysis to elucidate immunologic and pathologic associations with human natural killer cells and will serve as a valuable resource for future investigations into the immunogenomics and disease processes involving KIRs.

### Methods

#### Allele inference using expectation maximization

To infer allele probabilities from a set of read alignments to a database of KIR alleles, we use an



expectation-maximization algorithm to aggregate the alignment data into an initial set of allele probability “expectations,” which is then used to further weight the alignment data in order to refine our estimates of KIR allele probabilities. Thus, given a bootstrap of read alignments with 100% identity to at least one KIR allele, KIR-CLE’s EM algorithm iteratively generates probability estimates for each KIR reference allele. Let:

- $m$  = the total number of alleles of this KIR gene
- $n$  = the total number of reads in this bootstrap
- $A$  = the set of KIR alleles  $\{a_1, a_2, \dots, a_m\}$
- $R$  = the set of BAM reads  $\{r_1, r_2, \dots, r_n\}$
- $x_r$  = the number of alleles that read  $r$  aligns to with 100% identity
- $t$  = each time step of the expectation maximization algorithm
- $\alpha$  = a heuristically chosen convergence threshold

We first initialize an  $m \times n$  “alignment matrix”  $M$  to encode our read alignments:

$$M_{a,r}^{t=0} = \begin{cases} \frac{1}{x_r}, & r \text{ aligns to } a \\ 0, & r \text{ does not align to } a \end{cases} \quad (1)$$

Next, using  $M$ , we compute an initial expectation vector  $E^t$  representing our rudimentary estimate of each KIR allele’s probability in this sample:

$$E_a^t = \sum_{r \in R} M_{a,r}^t \quad (2)$$

Then, at each time step  $t$  of the expectation-maximization algorithm, we update the values of our alignment matrix  $M$  in a Bayesian fashion using the previously generated expectation vector  $E^t$  as our prior and  $M^t$  as our likelihood:

$$M_{a,r}^{t+1} = \frac{M_{a,r}^t * E_a^t}{\sum_a (M_{a,r}^t * E_a^t)} \quad (3)$$

Subsequently, we may generate an updated expectation vector  $E^{t+1}$  using  $M^{t+1}$  in conjunction with Eq. (2) above.

We continue to iterate through our expectation-maximization algorithm in this manner, computing  $E^t$  from  $M^t$  and then  $M^{t+1}$  from  $E^t$  and  $M^t$ , until we achieve our convergence criterion, defined as the sum of squared changes in  $M$  not exceeding a heuristically selected hyperparameter  $\alpha$ :

$$\sum_{a \in A, r \in R} (M_{a,r}^t - M_{a,r}^{t-1})^2 \leq \alpha \quad (4)$$

Ultimately, our final expectation vector  $E^T$  is outputted as our vector of allele probability estimates.

### Allele coding region collapse

Because we used WES data as our input, KIR variants that differed only at non-exonic sites were merged by summing their allele probability estimates. Furthermore, as we are primarily interested in the phenotypic effects of altered binding affinity to KIR domains, we merged variants that differ only by a silent mutation by summing their allele probability estimates as well. Thus, all KIR alleles subsequently are reported as a three-digit number following the KIR gene name (e.g., KIR2DL4\*005).

### KIR allele correlations with molecular and clinical correlates

Correlations between inferred KIR alleles and molecular and clinical correlates in TCGA and UK Biobank were performed by comparing the values of these correlates in samples with a given KIR allele versus those without that KIR allele. Comparisons were performed using two-sided Fisher’s exact tests for categorical variables and two-sided Mann-Whitney  $U$  tests for continuous variables. Correction for multiple hypothesis testing in all analyses was performed using the Bonferroni method.

### Clinical effect model comparison

To explore several correlations we discovered between KIR3DL2 genotype and the earliest age of atopy onset more deeply, we attempted to model the earliest age of atopy onset as a linear function of the KIR3DL2 genotype

$$(Age \text{ of Atopy Onset}) = m * x(genotype) + b \quad (5)$$

where  $x(genotype)$  is determined by the model of KIR allele expression. Under a dominant model, both homozygotes and heterozygotes for an allele KIR3DL2\*000 contribute equally to the phenotype. Thus,

$$x(genotype) = \begin{cases} \mathbf{1} & 2 \text{ copies of KIR3DL2} * 000 \\ \mathbf{1} & 1 \text{ copy of KIR3DL2} * 000 \\ \mathbf{0} & 0 \text{ copies of KIR3DL2} * 000 \end{cases} \quad (6)$$

Meanwhile, under a semi-dominant model, homozygotes are twice as expressive as heterozygotes:

$$x(genotype) = \begin{cases} \mathbf{2} & 2 \text{ copies of KIR3DL2} * 000 \\ \mathbf{1} & 1 \text{ copy of KIR3DL2} * 000 \\ \mathbf{0} & 0 \text{ copies of KIR3DL2} * 000 \end{cases} \quad (7)$$

Finally, under a recessive model, only homozygotes have an effect on phenotypic expression:

$$x(genotype) = \begin{cases} \mathbf{1} & 2 \text{ copies of KIR3DL2} * 000 \\ \mathbf{0} & 1 \text{ copy of KIR3DL2} * 000 \\ \mathbf{0} & 0 \text{ copies of KIR3DL2} * 000 \end{cases} \quad (8)$$

How well each of these models performed against each other was assessed by the goodness of fit of the final linear model (Eq. 5) with the UK Biobank data and summarized using the Bayesian Information Criterion.

### Software used

All analyses in this article were performed using an array of standard tools for bioinformatics. Pre-processing and processing of sequencing data as part of KIRCLE were accomplished using samtools 1.3.1, Nucleotide BLAST v2.8.1+, Python v3.6.10, argparse v1.1, numpy v1.18.1, pandas v1.0.1, and pysam v0.9.1. Later downstream statistical analyses and visualizations were performed using Python v3.7.7 with packages scipy v1.4.1, sklearn v0.22.1, matplotlib v3.1.3, and seaborn v0.10.1. Scripts were run on Red Hat Enterprise Linux Server release 7.4 (Maipo) on the BioHPC-Nucleus Supercomputer at UT Southwestern Medical Center.

### Abbreviations

NK: Natural killer; KIR: Killer immunoglobulin receptor; HLA: Human leukocyte antigen; NGS: Next-generation sequencing; KIRCLE: KIR CaLling by Exomes; WES: Whole exome sequencing; TCGA: The Cancer Genome Atlas; MSA: Multiple sequence alignment; NMDP: National Marrow Donor Program; FDR: False discovery rate; t-SNE: t-distributed Stochastic Neighbor Embedding; ICD10: International Statistical Classification of Diseases 10; PUD: Peptic ulcer disease; IQR: Interquartile range.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12915-022-01392-2>.

#### Additional file 1.

**Additional file 2: Supplementary Figure S1.** Validation of KIRCLE's Consistency. Contour plots demonstrating the effect of varying  $p$  and  $t$  on consistency of genotype calling, as quantified by entropy, for (a) KIR2DL2; (b) KIR2DL3, KIR2DL5B, KIR2DS2, & KIR3DS1; (c) KIR2DL5A; (d) KIR2DS1; (e) KIR2DS3; (f) KIR2DS4; (g) KIR2DS5; (h) KIR3DL1; (i) KIR3DL2; (j) and KIR3DL3. **Supplementary Figure S2.** Validation of KIRCLE's Accuracy. (a) Confusion matrix depicting KIRCLE's consistency in KIR genotype inference between 531 samples and their biological replicates in TCGA. (b) Scatterplot demonstrating that the concordance of KIR genotypes between TCGA replicates in each KIR gene positively correlates with sequence "dissimilarity" (as measured by average MSA distance) between alleles of that KIR gene. (c) Example MSA phylograms for KIR2DS2 (left) and KIR2DL5A (right) demonstrate different degrees of sequence similarity between different alleles of the same KIR gene. (d) Average entropy versus average depth of coverage in TCGA for each KIR gene. (e) Comparison of allele frequency ranks among the 9 KIR3DL2 alleles observed in the US NMDP with their frequency ranks in TCGA (left) and UK Biobank (right). **Supplementary Figure S3.** KIR Correlations with Molecular Markers. (a) Heatmap of the log<sub>2</sub>-odds-ratios of KIR allele correlations with HLA alleles in UK Biobank. Correlations with Fisher's Exact FDR > 0.25 were masked. (b) Volcano plot of KIR allele correlations with HLA alleles in TCGA. (c) Heatmap of log<sub>2</sub>-median-fold-changes in tumor immune infiltrate composition estimates stratified by KIR alleles in TCGA. Mann-Whitney-U FDR > 0.25 correlations were masked. (d) Volcano plot of KIR allele correlations with differences in tumor immune infiltrate composition in TCGA. (e) Heatmap of log<sub>2</sub>-median-fold-changes in other immune-related molecular signatures and markers stratified by KIR alleles in TCGA. Mann-Whitney-U FDR > 0.25 correlations were masked. (f) Volcano plot of KIR allele correlations with

differences in other immune-related molecular signatures and markers as measured and characterized in TCGA. **Supplementary Figure S4.** KIR Correlations with Other Clinical Variables in UK Biobank. (a) Increased likelihood of loose teeth was observed among individuals possessing at least one copy of the KIR3DL3\*002 allele compared to those without it. (b) Quantitative blood analysis of individuals homozygous for KIR3DL2\*010 revealed increased reticulocyte percentage compared to those without the allele. (c) Decreased waist circumference was observed in individuals possessing at least one copy of KIR2DL3\*010. (d) Lower age of cancer diagnosis was observed among KIR2DS3\*002 heterozygotes. (e) Lower age of Chronic Obstructive Pulmonary Disease (COPD) diagnosis was observed among KIR3DL2\*008 heterozygotes. (f) Increased duration of sleep was observed in individuals possessing at least one copy of KIR2DL4\*032. All error bars in bar plots depict standard error of mean.

**Additional file 3: Supplementary Table S1.** KIRCLE-inferred KIR allele genotypes in TCGA. Number of copies of each of 535 KIR coding alleles in 10,332 TCGA samples.

**Additional file 4: Supplementary Table S2.** KIRCLE-inferred KIR allele genotypes in UK Biobank. Number of copies of each of 535 KIR coding alleles in 49,694 UK Biobank samples. Note that UK Biobank ID numbers are study-dependent and not conserved between datasets. Therefore, if readers wish to make comparisons with correlates downloaded for their own project, they will need to re-compute these genotypes using KIRCLE on WES CRAM files that they themselves downloaded from UK Biobank.

### Acknowledgements

This research was supported in part by the computational resources provided by the BioHPC supercomputing facility located in the Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, TX. URL: <https://portal.biohpc.swmed.edu>.

This research has been conducted using the UK Biobank Resource under Application Number 21237. We are grateful for support from the UK Biobank Access Management Team for their assistance in accessing data from UK Biobank.

### Authors' contributions

B.L. proposed the initial idea of using exome data for KIR typing. G.F.G. and B.L. conceived and designed KIRCLE. G.F.G. performed the benchmarking analysis, as well as the association analysis between KIR alleles and molecular and clinical correlates. X.Z. helped with data access and management. G.F.G. and B.L. wrote the manuscript. D.L. and X.Z. offered suggestions and edited the manuscript. B.L. led the project. All authors read and approved the final manuscript.

### Funding

This work was funded in part by the American Federation for Aging Research (AFAR) and the National Institute on Aging (NIA) through the Medical Student Training in Aging Research (MSTAR) program (GG), Cancer Prevention and Research Institute of Texas (CPRIT) RR170079 (BL) and NCI grants 1R01CA245318-01 (BL) and 1R01CA258524-01 (BL). This work was also supported by the National Institutes of Health [5P30CA142543, 5R01GM126479, 5R01HG008983] and Cancer Prevention & Research Institute of Texas [CPRIT RP190107] (XZ).

### Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. TCGA datasets analyzed during the current study are available at the Genomic Data Commons of the National Cancer Institute, <https://gdc.cancer.gov/>. UK Biobank datasets analyzed during the current study are available from UK Biobank (<https://www.ukbiobank.ac.uk/>), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of UK Biobank. KIR allele calls in TCGA and UK Biobank generated in this manuscript have been included as Supplementary Tables 1 and 2 respectively. All other datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request. Code used to generate results central to this article's conclusions may be accessed as detailed below:

Project name: KIRCLE  
 Project home page: <https://github.com/gaog94/KIRCLE>  
 Archived Version: <https://zenodo.org/badge/latestdoi/189675746>  
 Operating system(s): Platform independent  
 Programming language: Python3  
 Other requirements: argparse v1.1, numpy v1.18.1, pandas v1.0.3, pysam v0.15.3, samtools 1.3.1, Nucleotide BLAST v2.8.1+  
 License: MIT  
 Any restrictions to use by non-academics: None

## Declarations

### Ethics approval and consent to participate

The need for Institutional Review Board Approval at our institution (University of Texas Southwestern Medical Center) was waived for this study as all data used for this project had previously been generated as part of either The Cancer Genome Atlas Project or UK Biobank and none of the results reported in this manuscript can be used to identify individual patients.

### Consent for publication

Not applicable

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>School of Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA. <sup>2</sup>Lyda Hill Department of Bioinformatics, University of Texas Southwestern Medical Center, Dallas, TX, USA. <sup>3</sup>Institute for Personalized Medicine, College of Medicine, Pennsylvania State University, Hershey, PA 17033, USA. <sup>4</sup>Department of Population and Data Sciences, University of Texas Southwestern Medical Center, Dallas, TX, USA.

Received: 8 September 2021 Accepted: 15 August 2022

Published online: 24 August 2022

## References

- Ong S, Rose NR, Cihakova D. Natural killer cells in inflammatory heart disease. *Clin Immunol*. 2017;175:26–33.
- O'Shea D, Hogan AE. Dysregulation of natural killer cells in obesity. *Cancers (Basel)*. 2019;11(4):573.
- Kamoda Y, Uematsu H, Yoshihara A, Miyazaki H, Senpuku H. Role of activated natural killer cells in oral diseases. *Jpn J Infect Dis*. 2008;61(6):469–74.
- Poggi A, Benelli R, Vene R, Costa D, Ferrari N, Tosetti F, et al. Human gut-associated natural killer cells in health and disease. *Front Immunol*. 2019;10:961.
- Hu W, Wang G, Huang D, Sui M, Xu Y. Cancer Immunotherapy based on natural killer cells: current progress and new opportunities. *Front Immunol*. 2019;10:1205.
- Marsh SG, Parham P, Dupont B, Geraghty DE, Trowsdale J, Middleton D, et al. Killer-cell immunoglobulin-like receptor (KIR) nomenclature report, 2002. *Immunogenetics*. 2003;55(4):220–6.
- Parham P. Immunogenetics of killer cell immunoglobulin-like receptors. *Mol Immunol*. 2005;42(4):459–62.
- Campbell KS, Purdy AK. Structure/function of human killer cell immunoglobulin-like receptors: lessons from polymorphisms, evolution, crystal structures and mutations. *Immunology*. 2011;132(3):315–25.
- Wagner I, Schefzyk D, Pruschke J, Schofl G, Schone B, Gruber N, et al. Allele-Level KIR Genotyping of More Than a Million Samples: Workflow, Algorithm, and Observations. *Front Immunol*. 2018;9:2843.
- Norman PJ, Hollenbach JA, Nemat-Gorgani N, Marin WM, Norberg SJ, Ashouri E, et al. Defining KIR and HLA class I genotypes at highest resolution via high-throughput sequencing. *Am J Hum Genet*. 2016;99(2):375–91.
- Kulkarni S, Martin MP, Carrington M. KIR genotyping by multiplex PCR-SSP. *Methods Mol Biol*. 2010;612:365–75.
- Tuttolomondo A, Di Raimondo D, Pecoraro R, Casuccio A, Di Bona D, Aiello A, et al. HLA and killer cell immunoglobulin-like receptor (KIRs) genotyping in patients with acute ischemic stroke. *J Neuroinflammation*. 2019;16(1):88.
- Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*. 2011;7:539.
- Gonzalez-Galarza FF, McCabe A, Santos E, Jones J, Takeshita L, Ortega-Rivera ND, et al. Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res*. 2020;48(D1):D783–D8.
- Gedil MA, Steiner NK, Hurley CK. KIR3DL2: diversity in a hematopoietic stem cell transplant population. *Tissue Antigens*. 2007;70(3):228–32.
- Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9:2579–605.
- Carrot-Zhang J, Chambwe N, Damrauer JS, Knijnenburg TA, Robertson AG, You C, et al. Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell*. 2020;37(5):639–54 e6.
- Norman PJ, Hollenbach JA, Nemat-Gorgani N, Guethlein LA, Hilton HG, Pando MJ, et al. Co-evolution of human leukocyte antigen (HLA) class I ligands with killer-cell immunoglobulin-like receptors (KIR) in a genetically diverse population of sub-Saharan Africans. *PLoS Genet*. 2013;9(10):e1003938.
- Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*. 2018;562(7726):203–9.
- Shukla SA, Rooney MS, Rajasagi M, Tiao G, Dixon PM, Lawrence MS, et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol*. 2015;33(11):1152–8.
- Long EO. Negative signaling by inhibitory receptors: the NK cell paradigm. *Immunol Rev*. 2008;224:70–84.
- Eltoukhi HM, Modi MN, Weston M, Armstrong AY, Stewart EA. The health disparities of uterine fibroid tumors for African American women: a public health issue. *Am J Obstet Gynecol*. 2014;210(3):194–9.
- Bellanti JA, Settignano RA. The atopic disorders and atopy ... "strange diseases" now better defined! *Allergy Asthma Proc*. 2017;38(4):241–2.
- Gragert L, Madbouly A, Freeman J, Maier M. Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry. *Hum Immunol*. 2013;74(10):1313–20.
- Oh S, Oh S. Epidemiological and genome-wide association study of gastritis or gastric ulcer in Korean populations. *Genomics Inform*. 2014;12(3):127–33.
- Lim YJ. Genetic susceptibility of gastroduodenal disease in ethnic and regional diversity. *Gut Liver*. 2014;8(6):575–6.
- Yun CH, Lundgren A, Azem J, Sjoling A, Holmgren J, Svennerholm AM, et al. Natural killer cells and *Helicobacter pylori* infection: bacterial antigens and interleukin-12 act synergistically to induce gamma interferon production. *Infect Immun*. 2005;73(3):1482–90.
- von Bubnoff D, Andres E, Hentges F, Bieber T, Michel T, Zimmer J. Natural killer cells in atopic and autoimmune diseases of the skin. *J Allergy Clin Immunol*. 2010;125(1):60–8.
- Deniz G, Akdis M. NK cell subsets and their role in allergy. *Expert Opin Biol Ther*. 2011;11(7):833–41.
- Khakoo SI, Carrington M. KIR and disease: a model system or system of models? *Immunol Rev*. 2006;214:186–201.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.