Cystatin C is a glucocorticoid response gene predictive of cancer immunotherapy failure

Sam O Kleeman[1], Miriam Ferrer[1], Breanna Demestichas[1], Sean Bankier[2,3], Hassal Lee[1], Todd Heywood[1], Arno Ruusalepp[4], Johan L. M. Bjorkegren[5], Brian R Walker[2,6], Hannah V Meyer[1,+], and Tobias Janowitz[1,7,+,*]

[1] Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA
[2] BHF Centre for Cardiovascular Science, Queen's Medical Research Institute, University of Edinburgh, Edinburgh, UK
[3] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway
[4] Department of Cardiac Surgery, Tartu University Hospital, Tartu, Estonia
[5] Department of Genetics & Genomic Sciences, Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6] Translational and Clinical Research Institute, Newcastle University, Newcastle upon Tyne, UK
[7] Cancer Institute, Northwell Health, New York, USA
[+] These authors jointly supervised this work
* Correspondence to janowitz@cshl.edu

Abstract

The secreted protein Cystatin C (CyC) is a cysteine protease inhibitor of incompletely characterized biomedical function, used clinically for estimation of glomerular filtration rate. Plasma CyC is elevated in many patients, especially when they receive glucocorticoid (GC) treatment. Here we empirically connect GCs with systemic regulation of CyC. First, we leveraged genome-wide association and structural equation modeling to determine the genetics of the latent trait CyC production in UK Biobank. Using multi-modal genomic, transcriptional, and experimental approaches, we demonstrated that CyC is a direct target of GC receptor, with GC-responsive CyC secretion exhibited by macrophages and cancer cells in vitro. Elevated serum CyC levels were positively correlated with GC levels in a murine model of cancer progression. Consistent with the coupling of CyC levels to GC signaling in a disease relevant manner, CyC predicted elevated all-cause and cancer-specific mortality in humans. These associations were orthogonally confirmed by a polygenic score (PGS) capturing germline predisposition to CyC production. This PGS predicted checkpoint immunotherapy failure in a combined clinical trial cohort of 685 metastatic cancer patients, with available germline exome sequencing. Taken together, our results demonstrate that CyC captures biomedically-relevant variations in endogenous GC activity, raising the possibility that CyC may be a direct effector of GC-induced immunosuppression and therefore a target for combination cancer immunotherapy.

Introduction

Large prospective patient cohorts with comprehensive genetic, physical and health data, often termed biobanks, allow for the re-evaluation of human disease, health and health care[1]. Due to substantial genetic variation between humans[2], genome-wide association studies (GWAS) are analogous to forward genetic screens, and can direct discovery of the molecular determinants of complex biomedically relevant phenotypes such as organ function[3]. Previously, we have developed a model for the accurate estimation of kidney filtration function, defined as the estimated glomerular filtration rate (eGFR), in patients with cancer[4,5.] Like others before, we used creatinine[6], a breakdown product of muscle creatine metabolism that is renally excreted[7], as a prediction variable. In non-cancer patients, this approach has been compared to the use of cystatin C (CyC)[8], a small secreted protein. Serum levels of both molecules depend on latent (unmeasured) components: most notably their production (reflecting both synthesis and externalization) from the producing cells and the GFR. While the determinants of creatinine production are relatively well-characterized and relate to muscle mass and diet[9], the factors that regulate CyC production are in contrast poorly understood[10]. Unlike the metabolic end-product creatinine, but like many other secreted proteins, cystatin C has biological functions: in its monomeric form it is a highly potent paracrine inhibitor of intracellular cysteine proteases[11,12].

Given the known function of CyC and its extracellular localization, it is likely that CyC production is systemically regulated. A first indication of this regulation may by the following set of observations. Organ transplant patients tend to have a higher serum CyC for a given measured GFR[13]. The vast majority of transplant patients are prescribed exogenous glucocorticoids (GCs), such as prednisolone or dexamethasone, as part of their immunosuppressive regimen[14]. Paired analyses accounting for patient-specific factors and renal function have demonstrated that exogenous GCs increase CyC production[15], an effect that has also been observed in patients with excess endogenous GC production (Cushing's syndrome)[16]. Moreover, CyC production is increased in a range of disease states that induce GC elevations, including viral infection[17], inflammatory disease[10] and cancer[18]. This positive association between GC exposure and CyC production has been recapitulated experimentally *in vitro*[19] and *in vivo*[20].

Cortisol, the endogenous GC in humans, is produced by the adrenal gland[21] in a circadian rhythm peaking in the early morning[22]. Through action on the cytosolic glucocorticoid receptor (GR), glucocorticoids profoundly modulate the cellular transcriptional landscape[23], affecting up to 20% of all genes[24] and driving systemic reprogramming of metabolism and immunity that is essential for life[25]. While our understanding of the mechanisms by which GCs are immune-modulatory remains limited[26], their immunosuppressive function is firmly established and therapeutically employed across a wide range of auto-immune and inflammatory diseases, such as rheumatoid

arthritis[27]. They are also used to mitigate immune-mediated damage to normal organ systems, a common and potentially severe side effect of T cell activation by checkpoint immunotherapy (CPI) in cancer[28]. This latter indication has emphasized the importance of determining whether, and in what circumstances, exogenous GCs could impair the efficacy of CPI[29,30]. Evidence from *in vivo* models of cancer suggest that even low doses of GCs can suppress anti-tumor immunity[30], leading to enhanced metastasis and reduced survival[31]. This has remained difficult to empirically investigate in cancer patients due to confounding by performance status and comorbidities[32], inconsistent CPI trial inclusion criteria[29] and the difficulties in performing well-controlled trials in this context.

While diurnal variation makes plasma cortisol an imprecise metric, there is clear evidence of inter-individual variation in cortisol production. Psychosocial stress[33], metabolic stress (for example, calorific restriction)[30], shift work[34] and chronic diseases, such HIV[35] and cancer[30], are associated with marked increases in endogenous GC production. This is not without consequence – alterations in cortisol production (and chronic GC treatment[36]) are associated with the onset of metabolic syndrome[37,38], a cluster of interrelated conditions including hypertension, insulin resistance, dyslipidemia and obesity, which is in turn associated with increased risk of cardiovascular disease and type 2 diabetes[39]. Twin studies have demonstrated relatively high heritability in plasma cortisol[40], suggesting that inter-individual variability in endogenous cortisol production may have a significant genetic component, which has remained largely unexplained by GWAS to date[41].

We hypothesized that, rather than being a passive marker of renal function, cystatin C is directly associated with disease states and that this association might be mediated by GC signaling. Here, to empirically investigate this question, we leverage UK Biobank (UKB), a prospective population-based cohort comprising approximately 480,000 subjects who provided germline genetics, serum CyC and serum creatinine. Using conventional GWAS for eGFR-CyC/eGFR-Creatinine followed by structural equation modeling (SEM), we estimate single nucleotide polymorphism (SNP)-level associations with the latent trait of CyC-production. We characterize patient-level predisposition to CyC-production via construction of a polygenic score (PGS), which is validated in a held-out cohort. Through colocalization analysis and gene set enrichment, coupled with validation through functional genomics and *in vitro* assays we demonstrate that CyC is a direct target of GR. We apply the CyC-production PGS to identify myeloid cells as the predominant source of CyC, which we confirm *in vitro*. Finally, using multi-omic analysis in several independent patient cohorts, we show that CyC-production is associated with metabolic syndrome, reduced cancer-specific survival and reduced efficacy of CPI.

Results

Genomic architecture of CyC production

To investigate the genomic architecture of CyC production, we first performed a discovery GWAS for eGFR-CyC and eGFR-Creatinine (eGFR-Cr) in 381,764 European subjects in UKB, using linear mixed models to account for population stratification and cryptic relatedness. We randomly selected 50,000 unrelated subjects from the overall UKB European population and excluded their data from the GWAS to enable later validation analyses (Figure 1b). Using linkage disequilibrium (LD) score regression, we identified strong genetic correlation (r^2=0.61) between eGFR-CyC and eGFR-Cr, consistent with both traits sharing a common factor that reflects renal filtration function. We reasoned that the genetic variance in eGFR-CyC that was not explained by this common factor represented the latent trait of CyC-production, given that the CyC plasma level is a function of both CyC excretion in the kidney and its cellular production. Thus, we estimated the single-nucleotide polymorphism (SNP)-level effects on CyC-production and renal function, by constructing a genomic structural equation model (SEM, Figure 1c, Figure S1a-b) implemented in Genomic-SEM[42], assuming no covariance between CyC-production and renal function. Providing confidence in our approach, loci known to directly regulate renal function such as *SHROOM3*[43] and *UMOD*[44] were predominantly associated with the renal function latent trait, while the locus coding for CyC (*CST3*) was predominantly associated with the CyC-production latent trait. Other loci associated with CyC-production, such as *SH2B3*[45] and *FLT3*[46], identify components of immune cell signaling cascades and are strongly associated with autoimmune disease. The index SNP at the *SH2B3* locus is a missense variant (R262W) and exhibits a markedly larger effect size than would be expected for its allele frequency (minor allele fraction = 0.48, Figure S1c), consistent with evidence that this variant is under active positive selection[47]. The *CPS1* locus, coding for carbamoyl-phosphate synthase 1, stood out as having divergent effects on renal function and CyC-production, probably reflecting its independent roles in creatine metabolism[48] and immune signaling[49]. We next performed tissue specific partitioned heritability analysis using gene expression and chromatin accessibility datasets and this confirmed enrichment of heritability of the renal function component in kidney tissues, in comparison to CyC-production (Figure S1d-e). This analysis also demonstrated enriched heritability for the renal function trait in liver tissues, which may reflect the coupling of hepatic and kidney function, observed clinically as hepatorenal syndrome[50].

Using the discovery data set, we captured the polygenic architecture of CyC-production by deriving a polygenic score (PGS), implemented in LDpred2[51] using HapMap3 variants, that could be reliably imputed in all of UKB, The Cancer Genome Atlas (TCGA) and Genotype-Tissue Expression (GTEx) cohorts (Figure S1f, Supplemental Data). We sought to maximize portability to

clinical sequencing cohorts where only exome sequencing is available, and thus derived a second PGS from HapMap3 variants that could be reliably imputed from exome sequencing data (Methods, Figure S1g). In order to validate both PGSs with the data from the validation set comprised of the 50,000 unrelated European patients (Figure 1b), it was necessary to define an independent patient-level estimate for CyC-production. This is possible, because the discordance between eGFR-CyC and eGFR-Cr can be viewed as an approximation for CyC-production. Therefore, we modelled eGFR-CyC as a function of eGFR-Cr and sex, and computed the residual (termed CyC-residual, Figure 1e). Using this CyC-residual as a surrogate for CyC-production, we confirmed that the genome-wide CyC-production PGS had significant predictive power in the validation cohort (r^2=0.08, p<1e-300, Figure 1f). As expected, predictive performance was reduced for the exome-wide PGS in the validation cohort (r^2=0.04, p<1e-300).

To investigate the trans-ancestral portability of the genome-wide CyC-production PGS, we measured performance versus CyC-residual in African (AFR, n=8152) and Central and South Asian (CSA, n=9845) genetic ancestry groups in UKB. We observed poor trans-ancestral portability of this PGS in these ancestry groups, with a low proportion of CyC-residual variance explained in CSA and AFR populations (Figure S2a-b). In order to derive a PGS in each non-EUR population, we performed GWAS and SEM as described above (Figure 1b) in these two ancestry groups but these analyses were under-powered to detect any signals reaching genome-wide significance (Figure S2c-d). While the genetic correlation between eGFR-CyC and eGFR-Cr in CSA subjects (r^2=0.65) was comparable to EUR subjects (r^2=0.61), genetic correlation was substantially diminished in AFR subjects (r^2=0.18). This indicates that eGFR-Cr and/or eGFR-Cy correlate weakly with true GFR in the AFR population, thus providing empirical genetic evidence to the observation that eGFR models have reduced performance in individuals self-identifying as Black or African American[52].

CyC is a glucocorticoid response gene

The *SERPINA1/6* locus on chromosome 14 had one of the largest effect sizes for CyC-production (Figure 1d, Figure 2a) and is known to be associated with plasma cortisol[41], suggesting a link between cortisol and CyC. In a recent cortisol Genome Wide Meta-Analysis (GWAMA), this signal was shown to be an expression qualitative trait locus (eQTL) for hepatic expression of *SERPINA6*[41], which encodes cortisol-binding globulin (CBG). To determine if there was a shared common variant, we performed co-localization analysis[53]. We did not detect a shared causal variant (posterior probability = 1.45e-15), but trans-eQTL analysis in the Stockholm Tartu Atherosclerosis Reverse Networks Engineering Task (STARNET)[54] cohort identified a single SNP (rs2749527) at the *SERPINA1/6* locus that was associated with significantly reduced plasma cortisol (p=1.75e-13) and significantly reduced *CST3* gene expression in a visceral adipose tissue

(p = 0.002448 in additive model, p=9.21e-6 in recessive model, Bonferroni-adjusted alpha level of 0.0025, Figure 2b). This variant is independently associated with significantly reduced liver *SERPINA6* expression in STARNET (p=4.73e-9, Figure 2c) and GTEx (p=0.004, Figure S3c) cohorts. As such, a single genetic instrument connects CBG, plasma cortisol and CyC (Figure S3a), thus providing genetic evidence for a direct link between GCs and CyC.

To further examine the link between GCs and CyC, we mapped each SNP meeting genome-wide significance to overlapping genes (defined by transcriptional start and end sites) and performed gene set enrichment analysis (GSEA) for gene sets relating to GC signaling. This analysis identified significant enrichment of 7/15 GC signaling gene sets from the Gene Ontology Resource (Figure 2d). In light of this, we hypothesized that *CST3* might be a direct transcriptional target of glucocorticoid receptor (GR, gene name *NR3C1*). Using functional genomics data derived from the ENCODE project, including chromatin immunoprecipitation sequencing (ChIP-seq) for GR and Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) data in the A549 cell line treated with dexamethasone, we identified dexamethasone-induced recruitment of GR to an accessible downstream enhancer element at the *CST3* locus (Figure 2e-f). In the same experiment, dexamethasone significantly increased *CST3* gene expression over time (p<0.0001, Figure 2g). We sought to investigate whether, and on what timescale, the transcriptional induction of CST3 by dexamethasone results in increased cellular secretion of CyC, which would cause increased tissue and circulating CyC levels. We first repeated the ENCODE experimental protocol using A549 cells and found that extracellular CyC concentration was significantly increased after 18 hours of treatment compared to 0 hours (Figure S3d). We next compared extracellular CyC concentration 18 hours after treatment with either dexamethasone or vehicle control in A549 cells (Figure 2h) and HeLa cells (Figure S3e) and found significant elevations of CyC in both cases. Consistent with this, subjects in UKB whose medical history included non-topical GC treatments had a significantly higher CyC-residual than GC-naïve subjects (p<1e-272, Figure 2i). We considered that there might be a direct or an indirect explanation for this clinical finding – either GC treatment increases CyC-production or an increased germline predisposition to CyC-production increases the risk for disorders that necessitate GC treatment (Figure S3f). To assess the dominant contribution, we compared CyC-production PGS between GC-treated and GC-naïve patients in the UKB validation set, and found only a marginal effect compared to the effect on CyC-production in the same cohort (t-statistic_{PGS} = -2.93, $\text{t-statistic}_{CyC\text{-}residual}$ = -10.8, Figure S3g). This suggests that the effect of GC treatment on CyC-production is dominant.

In light of our finding that extracellular CyC concentrations do not significantly increase until 18 hours after dexamethasone treatment *in vitro* (Figure S3d), we hypothesized that CyC would not show a circadian rhythm entrained by plasma cortisol levels. Despite the fact that plasma cortisol was not directly measured in UKB, we were able to assess this, because previous studies have

demonstrated that plasma bilirubin is strongly correlated to plasma cortisol over time[55]. Using cosinor regression[56] for CyC-residual and bilirubin in UKB subjects with available blood sampling, we determined that the diurnal variation of CyC-residual is diminished compared to bilirubin (amplitude coefficient = 0.038 versus 0.23, Figure S3h). Moreover, the peak in CyC-residual appears to be delayed by approximately 18 hours compared to the expected peak in plasma cortisol, consistent with our *in vitro* experiments. Taken together, these findings demonstrate that CyC production is directly induced by GCs, and that CyC has a reduced diurnal amplitude and offset periodicity compared to plasma cortisol.

CyC is produced by myeloid cells

The GC inducible component of CyC production indicates a potential biological relevance of CyC modulation, which would presumably operate in non-cancer cells. To assess this, we first determined whether CyC is secreted in a cell type-specific manner. On first glance, *CST3* gene expression was relatively consistent across all tissues examined as part of the GTEX project (GTEX Portal), but we reasoned that tissues that predominantly secrete CyC would exhibit a significant positive correlation between CyC-production PGS and *CST3* gene expression. Using expression qualitative trait score (eQTS) analysis, we detected a significant positive correlation in spleen tissues (n=171, Figure 3a). In support of this, we identified circadian rhythmicity from cosinor regression of spleen *CST3* gene expression against time of death, which was attenuated compared to the canonical GC target *FKBP5* (amplitude = 0.060 versus 0.24, Figure S4a). To understand which cell types might be driving this signal, we examined available single-cell RNA sequencing (scRNA-seq) data from human spleen[57]. This showed that only myeloid-derived cell populations (dendritic cells, macrophages and monocytes) expressed *CST3* (Figure 3b). We confirmed myeloid-specific *CST3* expression in peripheral blood mononuclear cells (PBMCs) scRNA-seq[58] (Figure 3c) and across multiple scRNA-seq datasets harmonized as part of the Human Protein Atlas[59] (Figure 3d). As additional validation, we found significant positive correlation between blood monocyte counts and CyC-residual in the UKB cohort (Figure S4b), while two-sample Mendelian randomization using blood-derived *CST3* eQTLs (eQTLGen[60]) as exposure identified a highly significant positive association with CyC-production (p=6.13e-77, Figure 3e). In light of evidence that GR is expressed at a substantially higher level in macrophages than monocytes[61], we hypothesized that monocytes might secrete CyC constitutively while macrophages would secrete CyC in response to GC signaling. To investigate this question, we used the monocyte-like THP-1 cell line which acquires macrophage-like differentiation upon treatment with the protein kinase C activator PMA. Dexamethasone treatment of macrophage-like THP-1 cells significantly increased *CST3* gene expression at 12 hours (Figure 3f) and extracellular CyC protein concentration at 18 hours (Figure 3g), mirroring the results found in A549 and HeLa cells. In contrast, there was no significant change in *CST3* expression (Figure S4c)

or extracellular CyC (Figure S4d) following dexamethasone treatment in monocyte-like THP-1 cells.

The findings that CyC is primarily expressed by myeloid cells and that GC-responsive CyC secretion is predominantly mediated by macrophages have the potential to explain our finding that rs2749527 is a trans-eQTL for *CST3* measured in visceral adipose tissue (VAT) in STARNET but not GTEX (p=0.77, Figure S3b). The STARNET study recruited patients with established coronary artery disease[54], while GTEX is a relatively unselected cohort of deceased donors[62]. As metabolic syndrome is associated with significant macrophage accumulation in adipose tissue[63], we hypothesized that STARNET patients would have significantly increased macrophage gene signatures in VAT, compared to GTEX donors. Using CIBERSORTx[64] (absolute mode) analysis of RNA-seq data in each cohort, we identified highly significant enrichment of M2 macrophages in STARNET versus GTEX (p=3.03e-289, Figure S5a). M2 macrophages were by far the most abundant myeloid component in the STARNET VAT samples, suggesting that they are the cell type underlying the trans-eQTL signal. This finding both provides orthogonal validation for the role of macrophages in GC-responsive CyC secretion and illustrates the limitations of eQTL analysis using bulk RNA-seq data, as has been described previously[65].

CyC is associated with diverse disease states

As increased endogenous GC production is a hallmark of disease progression and cortisol levels are known to be substantially elevated in post-mortem plasma samples[66], we hypothesized that CyC production would be associated with all-cause mortality in UKB. To investigate this, we first used multivariate Cox regression to estimate the effect of CyC-residual on all-cause mortality, adjusted for relevant patient covariates known to predict mortality[67–69]. We found that CyC-residual was associated with significantly increased all-cause mortality (HR=1.56, p<1e-16, Figure 4a). We considered that CyC-residual has the potential to be confounded by environmental factors, including exogenous GC treatment and, to mitigate this, we investigated whether the germline predisposition to CyC production, estimated as CyC-production PGS, could predict lifespan in our UKB European validation set (Figure 1b). Using multivariate Cox regression adjusted for sex, year of birth and genetic ancestry, we found that CyC-production PGS was associated with significantly reduced lifespan of UKB subjects (p=0.00013), as well as their two parents (p<1e-16, Figure 4b). Considering that the association between GC production and mortality is thought to be at least partly mediated via metabolic syndrome and resulting cardiovascular mortality[38], we hypothesized that CyC-production would be associated with accelerated onset of metabolic syndrome. To investigate this, we performed a phenome-wide association analysis (PheWAS) in the UKB validation set to identify time-to-event phenotypes (n=694) that were significantly associated with CyC-production using multivariate Cox regression

(Figure 4c). We identified positive associations meeting phenome-wide significance (p<1e-5) between CyC-production PGS and multiple diseases linked to metabolic syndrome, including type 2 diabetes, obesity, hypertension and ischemic heart disease, the latter finding consistent with a recent plasma cortisol Mendelian randomization analysis[41]. While we cannot exclude a GC-independent mechanism, the fact that CyC-production PGS recapitulates known GC-disease associations suggests that CyC-production captures sufficient inter-individual variation in endogenous GC production.

To extend these findings to specific disease states, we examined whether CyC-residual and CyC-production PGS were independent predictors of adverse outcomes in patients with cancer. Using UKB patients diagnosed with cancer since 2000 and with cancer-specific mortality defined by manual review of death certificates, we found that CyC-residual is an independent predictor of increased cancer-specific mortality (HR = 1.22, p<1e-16, Figure 4a). To validate this finding, we performed multivariate Cox regression of cancer-specific mortality against CyC-production across 13 tumor groups in 2 independent cohorts (UKB validation set, TCGA European subjects). Both fixed and random effect meta-analyses in each independent cohort confirmed a significant positive association between CyC-production PGS and cancer-specific mortality (Figure 4d-e). To reverse-translate these findings into an *in vivo* setting, we measured serum CyC concentrations in the murine colorectal-26 (C26) model of cancer progression. We identified significantly increased serum CyC during this cancer progression (Figure 4f). In the same *in vivo* model, we detected strong positive correlation between the murine endogenous GC, corticosterone, and CyC (Figure 4g). Considering that *in vitro* exogenous GC treatment of tumor cell lines A549 and HeLa (Figure 2h, Figure S3a) significantly increases CyC secretion, we hypothesized that *CST3* might be ectopically expressed in tumor cells. We re-analyzed scRNA-seq data from melanoma specimens from 12 patients[70], confirming high *CST3* expression in the myeloid compartment and identifying comparable ectopic *CST3* expression in the tumor compartment (Figure 4h). To confirm that this is a true emergent phenotype of tumor cells, rather than a reflection of increased intratumoral GC levels[26], we examined single-cell expression of the canonical GC target FKBP5, which was highly expressed in most intratumoral cell types (Figure S5b). Altogether, our findings suggest that the association of CyC with both cancer progression and cancer-specific survival may be mediated by GC-responsive ectopic expression of CST3 in tumor cells, which co-opt a pathway normally specific to macrophages.

CyC is associated with failure of cancer immunotherapy

As chronic exogenous GC treatment can suppress anti-tumor immunity[30] and CyC is induced by GC signaling, we hypothesized that CyC might directly alter the immune landscape of tumors. To investigate this question, we first explored the association between intratumoral *CST3* expression

and a set of pre-defined transcriptional immune markers[71] in the pan-cancer TCGA cohort. Using Spearman correlation analysis adjusted for multiple hypotheses, we found that *CST3* expression was significantly associated with reduced T- and B-cell clonality and IFNγ response, as well as increased M2 macrophage abundance (Figure 5a). Consistent with this, *CST3* gene expression was significantly elevated in the C5 TCGA immune subtype ($p<1e-10$, Tukey's test against all other subtypes), which is characterized by the highest macrophage and lowest lymphocyte abundance. This raises the question of whether increased *CST3* gene expression would be associated with resistance to CPI, a question that can best be answered from analysis of cell type-specific *CST3* expression. Therefore, we reviewed paired pre- and post-treatment tumor biopsy scRNA-seq from patients (n=8) with metastatic basal cell carcinoma (BCC), treated with anti-PD-1[72]. Patients were split into responders (n=3) and non-responders (n=5, Figure S6b) according to radiological response. From our data so far, it is clear that CyC is dynamically regulated and, consistent with this, pre-treatment *CST3* expression in macrophages, dendritic cells (DCs), cancer-associated fibroblasts (CAF) and tumor clusters did not predict CPI responsiveness ($p>0.5$, paired t-test). However, we observed evidence for significant dynamic *CST3* upregulation in CAFs and DCs in non-responder patients ($p<0.05$, paired t-test, Figure S3c-f).

We considered that the CyC-production PGS could reasonably capture the capacity to dynamically regulate secretion of CyC in response to GCs, and so predict failure of CPI. To estimate CyC-production PGS in patients treated with CPI, we collated 8 published cohorts of cancer patients treated with anti-PD-1, anti-PD-L1 or anti-CTLA-4 therapies with available germline exome sequencing (termed panIO cohort, Figure S6b, Table S1). 685 patients with European ancestry passed quality control for inclusion (cohort characteristics summarized in Table 1). Following imputation of common variants, the exome-wide CyC-production PGS was scored in each patient. Using multivariate Cox regression adjusted for sex, genetic ancestry and tumor type, we demonstrated that CyC-production PGS was associated with significantly worse progression-free survival (HR=1.29, p=0.00051) and worse overall survival (HR=1.09, p=0.10, Figure 5b). Using logistic regression with the same covariates, we further demonstrated that the PGS was associated with significantly reduced odds of durable clinical benefit (OR=0.78, p=0.0029, Figure 5b). This latter effect was broadly consistent in each tumor type. Altogether, these findings suggest that increased intratumoral CyC production, which in turn may be secondary to GC elevations, could make a significant contribution to failure of cancer immunotherapy.

Discussion

This work firmly establishes a direct link between GC signaling and CyC and we demonstrated its biological and clinical relevance using a combination of genetic analyses, *in vitro* and *in vivo* experiments, and clinically relevant prognostic and predictive studies. While the focus on human datasets allowed us to investigate clinically-relevant questions, we acknowledge that many of the analyses presented are limited by their associative nature. Although associations between measured CyC and clinical outcomes have the potential to be confounded in multiple directions, we agree with others that associations between patient-level polygenic scores and outcomes are substantially more robust, with the potential to capture causal associations[73]. In addition, we performed all PGS analyses in either a held-out validation cohort (for UKB) or an independent non-overlapping cohort (TCGA, panIO), to mitigate against the risk of overfitting. Thus, we argue that our findings make a material contribution to two gaps in our knowledge.

Firstly, studies of extreme GC biology have established causal associations with metabolic syndrome[38] and failure of cancer immunotherapy[31]. This causal link has remained elusive for intermediate GC perturbations, as circadian amplitude encompasses the inter-individual and disease-induced variation in cortisol for the majority of individuals. By demonstrating that CyC is a direct GC target with substantially reduced diurnal variability, we propose that CyC production can be used as a surrogate for GC signaling. With regard to immediate clinical application, our findings show that serum CyC does not merely have clinical utility as a marker of renal function, but should be considered as an exploratory biomarker to capture GC-associated disease. We recognize that while GR may not be the only transcription factor regulating CyC production and that we cannot be certain what proportion of variance in CyC-production we can explain on the basis of variation in cortisol. However, GC signaling it is likely to be a highly relevant regulator of CyC-production in the context of disease states. Through genome-wide association analysis for eGFR metrics derived from CyC and creatinine, coupled with structural equation modelling, we estimated the genetic contribution to the latent trait of CyC-production. Using a validated PGS derived from this trait, we demonstrate that CyC-production recapitulates known associations of GC signaling with all-cause mortality and onset of metabolic syndrome. The idea that germline-encoded GC tone could predispose to a cluster of 'preventable' diseases, including obesity, has significant implications to human medicine, and suggests the morbidity of subclinical hypercortisolism is underappreciated.

Secondly, despite widespread adoption of exogenous GCs as treatment for inflammatory conditions, the exact upstream mechanisms by which GCs cause immunosuppression remains elusive[26]. Our findings suggest that CyC should be viewed as a candidate effector of GC-induced immunosuppression. On GC stimulation, CyC is secreted in a cell-type specific manner by

macrophages, which would normally be recruited to tissues in the context of inflammation. Tumor-associated macrophages can directly synthesize endogenous GCs[26], suggesting that CyC secretion might be induced in an autocrine manner. Furthermore, CyC has the potential to directly suppress antigen presentation, a key upstream pathway in immune activation at the crossroads between innate and adaptive immunity. In its monomeric form, CyC is a potent inhibitor of cysteine proteases from the cathepsin (such as cathepsin B) and peptidase C13 (such as legumain) families, which are both directly involved in processing of lysosomal antigens for MHC class II presentation[12,74,75]. As extracellular CyC can be internalized by some epithelial cells at physiological concentrations[11], CyC can function as a paracrine regulator of lysosomal cysteine proteases. CyC levels are highest in cerebrospinal and seminal fluid[12], which may even suggest a role in immune privilege. Consistent with a direct immunosuppressive function of CyC, we demonstrate that germline predisposition to CyC production is significantly associated with substantial remodeling of the intertumoral immune landscape and failure of cancer immunotherapy. The evidence that CyC-production PGS predicts failure of immunotherapy requires experimental confirmation that is beyond the scope of this present study, but is supported by evidence that germline *CST3* knockout abrogates metastasis *in vivo*[76], a phenotype known to be immune-regulated[31]. If confirmed, it would suggest that combination PD-1/PD-L1 blockade and CyC inhibition should be explored as a putative therapeutic approach in patients who do not respond to CPI.

Methods

Cohort genomic data quality control (QC)

UK Biobank (UKB)

UKB-provided measured genotype, imputed genotype (GRCh37, imputed data release 3) and phenotype data[76] was accessed as part of application 58510. We selected subjects with available imputed genomic data (field 22028) and at least one paired creatinine (field 30700) and CyC measurement (field 30720), and excluded subjects with sex chromosome aneuploidy (field 22019), discordant genetic sex (fields 31 and 22001), excess heterozygosity and missing rate (field 22027). To classify genetic ancestry, we lifted over directly genotyped and linkage disequilibrium (LD)-pruned high-quality variants (bi-allelic SNPs, MAF > 0.1%, call rate > 99%) to GRCh38 and merged with variants available from an integrated callset (call rate > 95%) derived from 1000 Genomes and Human Genome Diversity Project (HGDP, gnomAD). LD pruning was implemented using PLINK1.9 with parameters '--indep-pairwise 50 5 0.2'. Principal components (1-10) were computed using the unrelated reference subjects (PC-relate kinship coefficient < 0.05) then projected onto all reference and UKB subjects. Next, a random forest classifier was trained using ancestry data from the reference cohort, implemented in the gnomAD package for Hail. This classifier was applied to the UKB subjects, and genetic ancestry was assigned with a minimum probability of 70% (Table S1). Relatedness data was extracted from the UKB-provided kinship matrix, generated using KING software. For the EUR ancestry group, subjects were split into a discovery cohort (n=381,764 subjects) and validation cohort (n=50,000 subjects), with the validation cohort comprising a random selection from unrelated UKB subjects (KING kinship coefficient < 0.0442). For all other ancestry groups, all subjects were used as discovery cohort. For all analyses using imputed data, we filtered to variants with INFO score > 0.8 and MAF > 1% across whole cohort.

Genotype-Tissue Expression (GTEx) project

Whole-genome sequencing data (GRCh38) and controlled-access metadata (including time of death) was accessed through dbGaP (phs000424.v8.p2) as part of application 26811. The provided imputed data had already undergone extensive quality control, however, we removed an additional 9 subjects with a PC-relate kinship coefficient > 0.05. We identified EUR ancestry subjects (n=678) as above using 1000G/HGDP reference data to train a random forest classifier that was applied to GTEX subjects, using high-quality LD-pruned common variants (bi-allelic SNPs, MAF > 0.1%, call rate > 99%, r^2<0.1), LD pruning was implemented using the 'ld_prune' function

in Hail, subsequent to removal of high-LD regions[31]. In this smaller cohort, ancestry was defined using a minimum probability of 50% followed by removal of PCA outliers with a PCA Z-score > 5.

The Cancer Genome Atlas (TCGA)

Germline array data (Birdseed format, GRCh37) was downloaded from the GDC Legacy archive as part of dbGaP application 26811, before conversion to VCF format. For sample QC, we started with a sample list defined by Sayaman et al[1], which selected one germline sample per subject, prioritizing blood-derived or high call-rate samples, while removing samples with excess heterozygosity or hematological malignancies. For additional sample QC, we removed samples with discordant sex (using the impute_sex function in Hail), excess hetero- or homo-zygosity (Z-score > 3, using agg.inbreeding function in Hail), related subjects (PC-relate kinship coefficient > 0.05) and called genetic ancestry as described for the GTEX cohort (n=7260 EUR patients). For imputation in the unrelated EUR population, we selected variants with call rate > 95% and MAF > 0.1%. Imputation was performed using the TOPMED server, which automatically lifts over variants to GRCh38. For the final cleaned dataset, we selected autosomal variants imputed with $r^2 > 0.6$ and MAF > 0.1%.

Stockholm Tartu Atherosclerosis Reverse Networks Engineering Task (STARNET)

STARNET is a cohort of 600 Caucasian patents of Eastern European origin, with a confirmed diagnosis of coronary artery disease. Genomic data quality control has been described previously[77]. Briefly, array-based genotyping was performed on germline DNA from blood, followed by imputation against the 1000 Genomes phase 1 SNPs. Comparison of population structure with 1000 Genomes cohort confirmed that all STARNET subjects had European genetic ancestry.

Immunotherapy meta-cohort (panIO)

We requested access to 8 cohorts of patients treated with CPI (anti-PD-1, anti-PD-L1 and/or anti-CTLA4) with available germline exome sequencing and clinical outcome (Table 1, Table S2). Clinical annotations were downloaded from the supplemental data from associated manuscripts or requested directly from principal investigators. Samples were excluded if there was insufficient data to report at least one outcome measure (overall survival, progression-free survival, durable clinical benefit). Durable clinical benefit (binary) was defined by patients with no radiological progression > 6 months or overall survival > 1 year. Harmonized germline short variant calling was implemented using nf-core/sarek pipeline, with Strelka mutation caller[78] and GRCh38 reference genome. gVCFs were merged using Illumina gvcfgenotyper tool and imported into Hail

for processing. Samples with discordant sex (n=13) were identified by comparison of sex reported in clinical metadata and genetic sex determined from integration of X chromosome heterozygosity and Y chromosome genotype counts (via PLINK 1.9 impute-sex function). For the small minority of patients without supplied sex (n=4), sex was genetically imputed. For variant QC, calls filtered by Strelka were removed, SNPs calls required a minimum depth of 7 while indel calls required a minimum depth of 10. Each variant required a call rate >90% and at least one 'high-quality' call defined as one homozygous ALT call or one heterozygous ALT call (with allele balance >15% for SNP or >20% for indel). Samples with a call rate <90% or excess hetero- or homozygosity (Z score > 3) were removed. No subjects had >3rd degree relatedness, which also excludes the possibility of duplicates samples in the cohort. EUR ancestry subjects were identified as for GTEX cohort. Imputation of EUR population was performed using the TOPMED server. For the final cleaned dataset, we selected autosomal variants imputed with $r^2 > 0.6$ and MAF > 0.1%.

Computation of principal components

For GWAS in UKB, we computed 20 principal components (PCs) on all subjects (including related) and all genotyped variants, as per the BOLT-LMM manual, implemented in PLINK2 (--pca function). Due to computational complexity, the PLINK2 PCA approximation (--approx) was used for the EUR population. To account for genetic ancestry in downstream analyses, PCs (1-4) were computed on high-quality linkage disequilibrium (LD)-pruned variants (bi-allelic SNPs, MAF > 0.1%, call rate > 99%, $r^2<0.1$), with SNPs in known high-LD regions removed[54]. For UKB, high-quality SNPs were derived from the 'in_PCA' field from the UKB-provided SNP QC file. In the UKB cohort, PCs were computed with related subjects removed (approach described above), and then projected onto all remaining samples, using the 'run_pca_with_relateds' function in the gnomAD package for Hail. In other cohorts (where related subjects were removed), PCs were computed using the 'hwe_normalized_pca' function in Hail.

Genome-wide association analysis (GWAS)

eGFR-CyC and eGFR-Cr were calculated using CKD-EPI equations implemented in the nephro package for R, with race term set to 0 for all subjects. For subjects we more than 1 paired creatinine and CyC measurement, we selected the earliest complete datapoint. Genome-wide association analyses (GWAS) in the discovery cohorts (for eGFR-CyC and eGFR-Cr were performed in each ancestry group, including related subjects, using BOLT-LMM[79] with covariates including age (field 21003), age^2, sex (field 31), genotyping array (binarized from field 22000), recruitment centre (field 54), and genetic PCs 1-20 (described above). LD score matrices for each ancestry group were downloaded from the Pan-UK Biobank project (https://pan.ukbb.broadinstitute.org/). To assess for confounding we determined the

attenuation ratio of each trait via LD score regression, which was within the expected range for polygenic traits (Table S3)[77].

Structural equation modelling

Structural equation modelling of eGFR-Cr and eGFR-Cy summary statistics was implemented in the Genomic-SEM package for R[80] and performed as per the GWAS-by-subtraction tutorial (https://rpubs.com/MichelNivard/565885). Briefly, for EUR, AFR and CSA populations, we performed LD score regression using LD matrices from the Pan-UK Biobank Project (https://pan.ukbb.broadinstitute.org/). We designed a structural equation model (summarized in Figure 1c), with latent traits estimated using the userGWAS function parallelized across each chromosome. Summary statistics for each latent trait (renal function, CyC-production) were extracted and effective sample sizes were estimated using the script provided by the Genomic-SEM authors (https://github.com/GenomicSEM/GenomicSEM/wiki/5.-User-Specified-Models-with-SNP-Effects). CyC-production summary statistics were standardized by setting A1 as the GRCh37 ALT allele and A2 as the GRCh37 REF allele, and multiplying the effect size of CyC-production by -1 so a higher effect size reflects increased CyC production.

Processing of summary statistics

Clumping was performed in the EUR eGFR-CyC summary statistics, implemented in PLINK 1.9 with parameters clump-r2 0.001, clump-p1 5e-8, clump-p2 5e-8 and clump-kb 10000 using 1000 Genomes reference data (derived from European subjects). For each clump, the index SNP (SNP with lowest p value) was annotated using the OpenTargets Genetics (https://genetics.opentargets.org/) variant-to-gene pipeline[80], which integrates both proximity and functional genomics data. For the small minority of variants (n=2) not represented in the OpenTargets database, the index SNP was annotated to the nearest coding gene. Partitioned heritability analysis was performed using the LDSC package for R[42] using the provided datasets, as per the tutorial by the package authors (https://github.com/bulik/ldsc/wiki/Cell-type-specific-analyses). For each trait and tissue-sample pair, we extracted the t-statistic as the ratio of the coefficient and standard error. To compare cell type-specific enrichment between renal function and CyC-production latent traits, we computed the absolute difference in t-statistic between eGFR-CyC and each latent trait, for each tissue sample. Colocalization analysis was performed using the coloc package for R[81], using the single-variant assumption. Gene set enrichment analysis of CyC-production latent trait was performed using MAGMA[82], implemented in the FUMA web server (https://fuma.ctglab.nl/) with a 0kb gene window. Mendelian randomization analysis, using cis-eQTLs probes for *CST3*, was implemented in GCTA-SMR[53] using SMR-formatted eQTL data from the eQTLGen Consortium[83].

Derivation and application of polygenic scores

CyC-production polygenic scores (PGS) were derived using LDpred2[84] (automatic model) according to the package vignette (https://privefl.github.io/bigsnpr/articles/LDpred2.html), using an initial estimated h^2 of 0.4. For the genome-wide score, HapMap3 variants were intersected with high-quality genomic variants available for all of UKB (array), TCGA (array) and GTEX (WGS) cohorts (n=1,031,527). For the exome-wide score, HapMap3 variants were intersected with high-quality exonic variants from the panIO cohort (n=352,549). The provided UKB LD reference was used for PGS derivation. Model fitting was confirmed by visual inspection of chain convergence for each PGS. The PLINK2 linear scoring function (--score) was used to apply the PGS to each cohort and to avoid exclusion of duplicate dbSNP IDs, the source data was filtered to the PGS variants according to position and alleles. The sample-level PGS was normalized by Z-scoring in each cohort. To generate a patient-level surrogate for CyC production, we modelled eGFR-CyC as a function of eGFR-Cr and sex, with intercept set as 0. We computed the residual of this model, termed CyC-residual, which is multiplied by -1 so that increasing CyC-residual reflects increased serum CyC relative to creatinine.

Functional genomics

ChIP-seq (for GR/NR3C1, timeseries accession: ENCSR210PYP) and ATAC-seq (timeseries accession: ENCSR385LRX) data for A549 cells treated with dexamethasone was downloaded from the ENCODE data portal (https://www.encodeproject.org/). Data was processed using the ENCODE data analysis pipeline, generating a p-value for each signal peak that reflects enrichment of DNA sequences. Data at the *CST3* locus was plotted using the karyoploteR package[60] for R. Enhancer activity scores ('ABC scores') derived from the validated activity-by-contact model[85], applied to 131 biosamples, was downloaded from ftp://ftp.broadinstitute.org/outgoing/lincRNA/ABC/Nasser2021-Full-ABC-Output/. Scores for the distal enhancer element at the *CST3* locus reflecting analysis of data derived THP-1 cells were extracted, and data from THP-1 cells treated with PMA was compared to data from naïve THP-1 cells.

Gene expression profiling

For GTEX and ENCODE gene expression profiling, gene-level counts derived from STAR-aligned RNA sequencing (RNA-seq) reads were downloaded from the GTEX (https://GTExportal.org/home/datasets) and ENCODE (timeseries accession: ENCSR897XFT) data portals respectively. TMM and library size normalization were applied using the edgeR package[51] for R, generating TMM-normalized log-counts per million (CPM) expression values that can be

compared between samples. For TCGA gene expression profiling, batch- and expression quantile-normalized data (RNA-seq) was downloaded from the PanCancer Atlas repository (https://gdc.cancer.gov/about-data/publications/pancanatlas). For STARNET gene expression profiling was performed as previously described[86] – briefly, gene counts were adjusted for GC content, library size and quantile-normalized implemented in EDAseq[87], prior to log-transformation. For digital cytometry analysis implemented in CIBERSORTx[54] (for STARNET and GTEX cohorts), gene expression was normalized to gene length to generate transcripts per million (TPM) expression values. CIBERSORTx was run in absolute mode with LM22 reference set, 100 permutations and B-mode batch correction.

Expression quantitative loci (eQTL) analysis

eQTLs were identified in the STARNET[88] cohort using the Kruskal Wallis test statistic (additive model), as implemented by the tool kruX[64], using individual-level genotype and gene expression data (data processing described above). To identify associations between *CST3* and SNPs present at the *SERPINA6/ SERPINA1* loci, we carried out this association analysis using 72 SNPs previously shown to be significantly associated with plasma cortisol[54]. This approach was applied to all non-vascular tissues (n=5) in the STARNET cohort (subcutaneous fat, visceral abdominal tissue, skeletal muscle, liver, blood). As the 72 SNPs reflecting 4 independent LD blocks, we modelled this analysis as 20 (4 LD blocks, 5 tissues) independent hypotheses and so the Bonferroni-corrected significance threshold was 0.0025. For independent validation of the significant eQTL associations in GTEX, we performed Kruskal Wallis tests in two tissues (visceral adipose tissue, liver) using the 'kruskal.test' function for R, using individual-level genotype and gene expression data (data processing described above). For further characterization of significant eQTLs, we constructed a recessive linear model of *CST3* gene expression as a function of genotype (binarized to 0/1 versus 2), using the 'lm' function for R.

Single-cell RNA sequencing (scRNA-seq) analysis

scRNA-seq expression matrices and metadata for Jerby-Anon et al.[70] and Yost et al.[72] were downloaded from Single Cell Portal (accession SCP109) and GEO (accession GSE123813) respectively. For Yost et al. the peritumoral T cell-specific samples were excluded from the analysis. Count or normalized expression data was imported into Seurat[89] (version 4.0), filtered (according to number of features, <10,000, and mitochondrial content, <7.5%, per cell), log-normalized (if applicable) and scaled. Highly variable features (n=2000) were used for principal component analysis followed by clustering (Louvain algorithm). Immune clusters were annotated by comparison to reference PBMC data, implemented in clustifyR package for R[70]. Unannotated clusters (presumed to reflect one of tumor, cancer-associated fibroblast or endothelial cells)

were manually annotated via established marker gene expression[72] and clonal copy number variation profiles, examined using the inferCNV[90] package for R. Patient-level pseudobulk cluster-specific expression data was extracted using the 'AverageExpression' function in Seurat.

Cosinor regression

Cosinor regression for blood markers (bilirubin, CyC-residual) and gene expression (*FKBP5*, *CST3*) as a function of time was performed using the cosinor package for R. A cosinor model has 4 parameters – MESOR (intercept), period (assumed as 24 hours), amplitude and acrophase (timing of activity peak). Bilirubin was extracted from UKB field 30840 and CyC-residual was derived as above; both metrics were standardized with Z-scoring stratified by age (decade) and sex. Gene expression was derived from TMM-normalized TPM data (implemented in edgeR) to facilitate intra- and inter-sample comparisons. For UKB, time referred to time of sampling and for GTEX, time referred to time of death; both rounded to the nearest hour in 24-hour clock. Amplitude coefficients were extracted from the transformed coefficients table.

UK Biobank cancer cohort

To identify patients who were treated with non-topical exogenous GCs, we reviewed field 20003 for coded medications bio-equivalent to dexamethasone or prednisolone. Subject lifespan was extracted from analysis of fields 40007 and 34. Parental lifespan was extracted from analysis of fields 2946, 1845, 1797, 1835, 1807 and 3526. Using cancer registry data (fields 40005, 40012, 40008, 40011), ICD10-coded cancer diagnoses were extracted and mapped to Phecodes (https://phewascatalog.org/). Using a curated list of operation codes (OPCS-4) reflecting curative procedures for 13 main tumor groups (Table S4), we mapped each cancer diagnosis to matched surgeries that occurred no more than 90 days prior to the coding entry. To account for variation in operation data availability prior to 2000, we filtered the data to cancers that were diagnosed after the year 2000. In cases where a patient was coded with a cancer of the same primary type more than once, the entries were merged. Patients with more than one discrete cancer diagnosis were excluded (n=2435 subjects) due to the difficulty in defining the time since diagnosis. For recruited patients who had died, we manually reviewed details from the death certificate (field 40010) to identify descriptions that were consistent with cancer-specific mortality.

Survival analyses

For Cox regression of overall and cancer-specific survival against CyC-residual, the time variable used was time from blood sampling to death or last follow-up date (nominally June 2020). For subjects with multiple CyC-residual datapoints over time, each datapoint was annotated with

survival time relative to blood sampling and treated independently. Model covariates included age (at blood sampling), sex, body mass index (BMI), hemoglobin, eGFR-Cr, C-reactive protein. For Cox regression of lifespan (for subject and parents) against CyC-production PGS in the UKB validation cohort, we used age at death or age at most recent follow-up as the time variable. Model covariates included year of birth (of subject, as parental birth years are not recorded) to account for historical increases in mean lifespan.

For Cox regression of cancer-specific survival against CyC-production PGS, it was necessary to consider bias from left truncation, where patients who died between diagnosis and the recruitment period would not be recruited. To account for this, the time interval used for Cox regression of overall and cancer-specific survival against CyC-production PGS in UKB referred to time from recruitment to death or data cut-off (June 2020). In contrast, TCGA patients were generally recruited close to the time of cancer diagnosis, prior to surgical resection of tumor and so time from diagnosis to death or last-follow-up date was used. Cancer-specific survival was extracted from the 'DSS' and 'DSS.time' fields in the TCGA clinical data resource available as part of the PanCancer Atlas (https://gdc.cancer.gov/about-data/publications/pancanatlas). Cancer-specific survival analyses with respect to CyC-production PGS were adjusted for age of diagnosis, genetic ancestry (PC1-4), sex (except sex-specific cancers), and a term reflecting whether curative surgery was performed. For UKB this term was derived from matching with curative operation codes as described above, for TCGA this term was derived from the field 'residual_tumor' in the clinical data resource. UKB-specific PGS-cancer survival analyses were additionally adjusted for recruitment center (to account for regional heterogeneity in cancer outcomes) and genotyping array. Pan-cancer inverse variance-weighted meta-analysis in each cohort (UKB, TCGA) was implemented in the meta package for R[91] using both fixed and random effects models.

For phenome-wide time-to-event analysis in UKB, all UKB 'first occurrence' fields and cancer registry data (fields 40005 and 40006) were extracted, with ICD10 codes mapped to Phecodes. If multiple ICD10 codes mapped to a single Phecode, the earliest date of diagnosis was selected. For each time-to-event Phecode, the time variable was defined as time from birth to first occurrence of diagnosis or most recent follow-up date. To account for region-specific variability in health record linkage, this date was determined by either the most recent coded diagnosis or most recent UKB center visit. Each phenotype-specific Cox regression was adjusted for sex, genetic ancestry (PC1-4) and year of birth (to account for historical variation in disease risk).

***In vitro* experiments**

Human lung carcinoma cell line A549 was purchased from ATCC (CCL-185). Human cervical cancer cell line HeLa was obtained from Cold Spring Harbor Laboratory. Human acute monocytic

leukemic cell line THP-1 was purchased from ATCC (TIB-202). A549 and HeLa cell lines were cultured in DMEM supplemented with 10% FBS and 1% penicillin-streptomycin. THP-1 cells were cultured in RPMI supplemented with 10% FBS and 1% penicillin-streptomycin. Cell viability was checked by trypan blue method and was consistently above 95% prior to seeding. For all experiments, cells were plated in 6 well plates, at a density of approximately 500,000 cells/ml. Cells reached confluence on day one or day two after being seeded. Macrophage-like differentiation in THP-1 cells was induced by treatment with 50nM PMA (Sigma) for 48 hours, before replacement with PMA-free media and recovery for 24 hours prior to treatment. For time course experiments, cells were seeded and harvested at the same time, with the only variable being the duration of treatment with 100nM dexamethasone (varied between 0 and 18 hours), with 0-hour treatment acting as the control. For single-timepoint experiments, cells were treated with either 100nM dexamethasone (Sigma) or 0.01% ethanol (vehicle) for 18 hours prior to harvesting. For each experiment all samples were harvested concurrently.

For quantification of extracellular CyC, cell supernatant was collected at harvesting, spun at 10000 x g for 5 minutes to remove debris, and analyzed by ELISA (Human Cystatin C ELISA Kit, R&D Systems), with each sample profiled in duplicate. For quantification of cellular protein content, cells were washed with DPBS and ice-cold RIPA buffer with protease and phosphatase inhibitors (Thermo Fisher) was added to each well. The cell lysate was passed through a 25G syringe for homogenization and spun for 10000 x g for 15 minutes, at 4°C. The supernatant from the spun-down lysate was then stored at -80 for later analysis. BCA assay was performed on the cell lysate, with each sample profiled in duplicate. Normalized extracellular CyC concentrations were determined by dividing the ELISA-derived CyC concentration by the BCA-derived cellular protein content.

For quantitative real-time polymerase chain reaction (RT-PCR), RNA was extracted using the RNeasy Mini Kit (Qiagen) and reverse transcribed using SuperScript IV VILO Master Mix (Thermo Fisher) according to the manufacturer's protocol. Four housekeeping genes (GUSB, PPIA, RPL15, RPL19) with minimal variation on GC treatment were selected on the basis of a literature review[92] and differential expression analysis in ENCODE RNA-seq data (accession ENCSR897XFT), implemented in edgeR. Primers were designed using NCBI Primer-BLAST, with exon-spanning primers designed where possible (primer sequences detailed in Table S5). PCR was performed using the PowerTrack SYBR Green Master Mix (Thermo Fisher) using the QuantStudio 6 Flex (Thermo Fisher) instrument, using a 10µl reaction volume in technical triplicate according to the manufacturer's protocol. The threshold cycle was determined by the Second Derivative Maximum method and the expression of each target was normalized relative to the geometric mean of four endogenous controls.

In vivo experiments

Experiments with the C26 model were performed using 8-weeks old wild-type BALB/c male mice obtained from Charles River Laboratories. Mice were allowed to acclimatize for 7 days and then were inoculated subcutaneously in their right flank with the syngeneic C26 colorectal cancer cell line ($2x10^6$ viable cells in 100µl RPMI vehicle) that induces cachexia. Mice were kept in pathogen-free conditions on a 24 hour 12:12 light-dark cycle. C26-tumor bearing mice were termed pre-cachectic from 18 days post-inoculation and were defined as cachectic when their weight loss exceeded 15% from peak body weight. All animal experiments and care were performed in accordance with the Cold Spring Harbor Laboratory (CSHL) Institutional Animal Care and Use Committee (IACUC) and the National Institutes of Health Guide for the Care and Use of Laboratory Animals. Plasma samples were obtained from tail bleeds and terminal cardiac bleeds. Tail bleeds were performed using a scalpel via tail venesection without restraint, and terminal bleeds were obtained at endpoint (cachexia) through exsanguination via cardiac puncture under isoflurane anesthesia. Samples were kept on ice at all times. Plasma samples were collected into heparin-coated capillary tubes to avoid coagulation and were processed as follows: centrifuge spin at 14,000 rpm for 5 minutes at 4°C, snap frozen in liquid nitrogen, and stored at -80°C. Plasma levels of CyC were determined with Mouse Cystatin C ELISA Kit (ab119590), Abcam. Corticosterone levels were quantified using Corticosterone ELISA (RE52211) from IBL International (TECAN). Prior to subcutaneous injection, C26 cells were cultured in RPMI 1640 Medium (+L-Glutamine) with 10% heat-inactivated FBS under sterile conditions. This was followed by Trypsin-enzymatic dissociation, resuspension in FBS-free RPMI, counting of viable cell concentration and injection of $2x10^6$ viable cells subcutaneously into the right flank of each mouse, in 100µl RPMI vehicle.

Statistical analysis

Significance testing refers to two-tailed unpaired t-tests with the assumption of unequal variance unless stated otherwise. For statistical and computational analyses, we used R (version 4.0.2) and Python (version 3.7.4) implemented as a Jupyter Notebook.

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* (2018) doi:10.1038/s41586-018-0579-z.
2. Auton, A. *et al.* A global reference for human genetic variation. *Nature* (2015) doi:10.1038/nature15393.
3. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *American Journal of Human Genetics* (2017) doi:10.1016/j.ajhg.2017.06.005.
4. Janowitz, T. *et al.* New model for estimating glomerular filtration rate in patients with cancer. *J. Clin. Oncol.* (2017) doi:10.1200/JCO.2017.72.7578.
5. Williams, E. H. *et al.* CamGFR v2: a new model for estimating the glomerular filtration rate from standardized or non-standardized creatinine in patients with cancer. *Clin. Cancer Res.* (2020) doi:10.1158/1078-0432.CCR-20-3201.
6. Rehberg, P. B. Studies on Kidney Function. *Biochem. J.* (1926) doi:10.1042/bj0200447.
7. Stevens, L. A., Coresh, J., Greene, T. & Levey, A. S. Assessing Kidney Function — Measured and Estimated Glomerular Filtration Rate. *N. Engl. J. Med.* (2006) doi:10.1056/nejmra054415.
8. BUTLER, E. A. & FLYNN, F. V. The occurrence of post-gamma protein in urine: a new protein abnormality. *J. Clin. Pathol.* (1961) doi:10.1136/jcp.14.2.172.
9. Levey, A. S. Measurement of renal function in chronic renal disease. *Kidney Int.* (1990) doi:10.1038/ki.1990.182.
10. Stevens, L. A. *et al.* Factors other than glomerular filtration rate affect serum cystatin C levels. *Kidney Int.* (2009) doi:10.1038/ki.2008.638.
11. Ekström, U. *et al.* Internalization of cystatin C in human cell lines. *FEBS J.* (2008) doi:10.1111/j.1742-4658.2008.06600.x.
12. Barrett, A. J., Davies, M. E. & Grubb, A. The place of human γ-trace (cystatin C) amongst the cysteine proteinase inhibitors. *Biochem. Biophys. Res. Commun.* (1984) doi:10.1016/0006-291X(84)91302-0.
13. Bökenkamp, A. *et al.* Cystatin C serum concentrations underestimate glomerular filtration rate in renal transplant recipients. *Clin. Chem.* (1999) doi:10.1093/clinchem/45.10.1866.
14. STARZL, T. E., MARCHIORO, T. L. & WADDELL, W. R. THE REVERSAL OF REJECTION IN HUMAN RENAL HOMOGRAFTS WITH SUBSEQUENT DEVELOPMENT OF HOMOGRAFT TOLERANCE. *Surg. Gynecol. Obstet.* (1963).
15. Risch, L., Herklotz, R., Blumberg, A. & Huber, A. R. Effects of glucocorticoid immunosuppression on serum cystatin C concentrations in renal transplant patients. *Clin. Chem.* (2001) doi:10.1093/clinchem/47.11.2055.
16. Naka, M. *et al.* Overestimation of glomerular filtration rate calculated from serum creatinine as compared with cystatin c in patients with subclinical hypercortisolism: Hyogo adrenal metabolic registry. *Endocr. J.* (2020) doi:10.1507/endocrj.EJ19-0478.
17. Odden, M. C. *et al.* Cystatin C level as a marker of kidney function in human immunodeficiency virus infection: The FRAM study. *Arch. Intern. Med.* (2007) doi:10.1001/archinte.167.20.2213.

18. Jones, M., Denieffe, S., Griffin, C., Tinago, W. & Fitzgibbon, M. C. Evaluation of cystatin C in malignancy and comparability of estimates of GFR in oncology patients. *Pract. Lab. Med.* (2017) doi:10.1016/j.plabm.2017.05.005.
19. Bjarnadóttir, M., Grubb, A. & Ólafsson, Í. Promoter-mediated, dexamethasone-induced increase in cystatin c production by hela cells. *Scand. J. Clin. Lab. Invest.* (1995) doi:10.3109/00365519509110261.
20. Zhu, X. R., Ge, N., Wang, Y., Zhai, J. L. & Liu, C. Corticosteroids significantly increase cystatin C levels in the plasma by promoting cystatin C production in rats. *Ren. Fail.* (2019) doi:10.1080/0886022X.2019.1638798.
21. Mason HL, Hoehn WM, K. E. Chemical studies of supra-renal cortex iv structure of compounds C-D-E-F and G. *J. Biol. Chem.* (1938).
22. Weitzman, E. D. *et al.* Twenty-four hour pattern of the episodic secretion of cortisol in normal subjects. *J. Clin. Endocrinol. Metab.* (1971) doi:10.1210/jcem-33-1-14.
23. Reddy, T. E. *et al.* Genomic determination of the glucocorticoid response reveals unexpected mechanisms of gene regulation. *Genome Res.* (2009) doi:10.1101/gr.097022.109.
24. Galon, J. *et al.* Gene profiling reveals unknown enhancing and suppressive actions of glucocorticoids on immune cells. *FASEB J.* (2002) doi:10.1096/fj.01-0245com.
25. Oelkers, W. Current concepts: Adrenal insufficiency. *N. Engl. J. Med.* (1996) doi:10.1056/NEJM199610173351607.
26. Acharya, N. *et al.* Endogenous Glucocorticoid Signaling Regulates CD8+ T Cell Differentiation and Development of Dysfunction in the Tumor Microenvironment. *Immunity* (2020) doi:10.1016/j.immuni.2020.08.005.
27. Hench, P. S. & Kendall, E. C. The effect of a hormone of the adrenal cortex (17-hydroxy-11-dehydrocorticosterone; compound E) and of pituitary adrenocorticotropic hormone on rheumatoid arthritis. *Proc. Staff Meet. Mayo Clin.* (1949).
28. Brahmer, J. R. *et al.* Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: American society of clinical oncology clinical practice guideline. *J. Clin. Oncol.* (2018) doi:10.1200/JCO.2017.77.6385.
29. Connell, C. M. *et al.* Cancer immunotherapy trial registrations increase exponentially but chronic immunosuppressive glucocorticoid therapy may compromise outcomes. *Annals of oncology : official journal of the European Society for Medical Oncology* (2017) doi:10.1093/annonc/mdx181.
30. Flint, T. R. *et al.* Tumor-Induced IL-6 Reprograms Host Metabolism to Suppress Anti-tumor Immunity. *Cell Metab.* (2016) doi:10.1016/j.cmet.2016.10.010.
31. Obradović, M. M. S. *et al.* Glucocorticoids promote breast cancer metastasis. *Nature* (2019) doi:10.1038/s41586-019-1019-4.
32. Janowitz, T., Kleeman, S. & Vonderheide, R. H. Reconsidering Dexamethasone for Antiemesis when Combining Chemotherapy and Immunotherapy. *Oncologist* (2021) doi:10.1002/onco.13680.
33. Michaud, K., Matheson, K., Kelly, O. & Anisman, H. Impact of stressors in a natural context on release of cortisol in healthy adult humans: A meta-analysis. *Stress* (2008) doi:10.1080/10253890701727874.
34. Manenschijn, L., Van Kruysbergen, R. G. P. M., De Jong, F. H., Koper, J. W. & Van Rossum,

E. F. C. Shift work at young age is associated with elevated long-term cortisol levels and body mass index. *J. Clin. Endocrinol. Metab.* (2011) doi:10.1210/jc.2011-1551.
35. Villette, J. M. *et al.* Circadian variations in plasma levels of hypophyseal, adrenocortical and testicular hormones in men infected with human immunodeficiency virus. *J. Clin. Endocrinol. Metab.* (1990) doi:10.1210/jcem-70-3-572.
36. Souverein, P. C. *et al.* Use of oral glucocorticoids and risk of cardiovascular and cerebrovascular disease in a population based case-control study. *Heart* (2004) doi:10.1136/hrt.2003.020180.
37. Mårin, P. *et al.* Cortisol secretion in relation to body fat distribution in obese premenopausal women. *Metabolism* (1992) doi:10.1016/0026-0495(92)90171-6.
38. Walker, B. R. Cortisol - Cause and cure for metabolic syndrome? *Diabet. Med.* (2006) doi:10.1111/j.1464-5491.2006.01998.x.
39. Eckel, R. H., Grundy, S. M. & Zimmet, P. Z. The metabolic syndrome. in *Lancet* (2005). doi:10.1016/S0140-6736(05)66378-7.
40. Bartels, M., Van den Berg, M., Sluyter, F., Boomsma, D. I. & De Geus, E. J. C. Heritability of cortisol levels: Review and simultaneous analysis of twin studies. *Psychoneuroendocrinology* (2003) doi:10.1016/S0306-4530(02)00003-3.
41. Crawford, A. A. *et al.* Variation in the SERPINA6/SERPINA1 locus alters morning plasma cortisol, hepatic corticosteroid binding globulin expression, gene expression in peripheral tissues, and risk of cardiovascular disease. *J. Hum. Genet.* (2021) doi:10.1038/s10038-020-00895-6.
42. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat. Hum. Behav.* (2019) doi:10.1038/s41562-019-0566-x.
43. Yeo, N. C. *et al.* Shroom3 contributes to the maintenance of the glomerular filtration barrier integrity. *Genome Res.* (2015) doi:10.1101/gr.182881.114.
44. Hart, T. C. *et al.* Mutations of the UMOD gene are responsible for medullary cystic kidney disease 2 and familial juvenile hyperuricaemic nephropathy. *J. Med. Genet.* (2002) doi:10.1136/jmg.39.12.882.
45. Smyth, D. J. *et al.* Shared and Distinct Genetic Variants in Type 1 Diabetes and Celiac Disease. *N. Engl. J. Med.* (2008) doi:10.1056/nejmoa0807917.
46. Saevarsdottir, S. *et al.* FLT3 stop mutation increases FLT3 ligand level and risk of autoimmune thyroid disease. *Nature* (2020) doi:10.1038/s41586-020-2436-0.
47. Zhernakova, A. *et al.* Evolutionary and Functional Analysis of Celiac Risk Loci Reveals SH2B3 as a Protective Factor against Bacterial Infection. *Am. J. Hum. Genet.* (2010) doi:10.1016/j.ajhg.2010.05.004.
48. Shin, S. Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* (2014) doi:10.1038/ng.2982.
49. Park, M. J. *et al.* Constitutive release of CPS1 in bile and its role as a protective cytokine during acute liver injury. *Proc. Natl. Acad. Sci. U. S. A.* (2019) doi:10.1073/pnas.1822173116.
50. Ginès, P. *et al.* Hepatorenal syndrome. *Nat. Rev. Dis. Prim.* **4**, 23 (2018).
51. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: Better, faster, stronger. *Bioinformatics* (2020) doi:10.1093/bioinformatics/btaa1029.

52. Diao, J. A. *et al.* In Search of a Better Equation — Performance and Equity in Estimates of Kidney Function. *N. Engl. J. Med.* (2021) doi:10.1056/nejmp2028243.
53. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet.* (2014) doi:10.1371/journal.pgen.1004383.
54. Franzén, O. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science (80-.).* (2016) doi:10.1126/science.aad6970.
55. Gu, F. *et al.* Diurnal variation of metabolites in three individual participants. *Chronobiol. Int.* (2019) doi:10.1080/07420528.2018.1541901.
56. Tong, Y. L. Parameter Estimation in Studying Circadian Rhythms. *Biometrics* (1976) doi:10.2307/2529340.
57. Madissoon, E. *et al.* ScRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol.* (2019) doi:10.1186/s13059-019-1906-x.
58. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* (2020) doi:10.1038/s41591-020-0944-y.
59. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* (2015) doi:10.1126/science.1260419.
60. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* (2018) doi:10.1101/447367.
61. Diaz-Jimenez, D., Petrillo, M. G., Busada, J. T., Hermoso, M. A. & Cidlowski, J. A. Glucocorticoids mobilize macrophages by transcriptionally up-regulating the exopeptidase DPP4. *J. Biol. Chem.* (2020) doi:10.1074/jbc.RA119.010894.
62. Carithers, L. J. & Moore, H. M. The Genotype-Tissue Expression (GTEx) Project. *Biopreservation and Biobanking* (2015) doi:10.1089/bio.2015.29031.hmm.
63. Weisberg, S. P. *et al.* Obesity is associated with macrophage accumulation in adipose tissue. *J. Clin. Invest.* (2003) doi:10.1172/jci19246.
64. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* (2019) doi:10.1038/s41587-019-0114-2.
65. van der Wijst, M. G. P. *et al.* The single-cell eQTLGen consortium. *Elife* (2020) doi:10.7554/eLife.52155.
66. Swaab, D. F. *et al.* Increased Cortisol Levels in Aging and Alzheimer's Disease in Postmortem Cerebrospinal Fluid. *J. Neuroendocrinol.* (1994) doi:10.1111/j.1365-2826.1994.tb00635.x.
67. Kaptoge, S. *et al.* C-reactive protein concentration and risk of coronary heart disease, stroke, and mortality: An individual participant meta-analysis. *Lancet* (2010) doi:10.1016/S0140-6736(09)61717-7.
68. Ganna, A. & Ingelsson, E. 5 year mortality predictors in 498 103 UK Biobank participants: A prospective population-based study. *Lancet* (2015) doi:10.1016/S0140-6736(15)60175-1.
69. Lees, J. S. *et al.* Glomerular filtration rate by differing measures, albuminuria and prediction of cardiovascular disease, mortality and end-stage kidney disease. *Nat. Med.* (2019) doi:10.1038/s41591-019-0627-8.

70. Jerby-Arnon, L. *et al.* A Cancer Cell Program Promotes T Cell Exclusion and Resistance to Checkpoint Blockade. *Cell* (2018) doi:10.1016/j.cell.2018.09.006.
71. Thorsson, V. *et al.* The Immune Landscape of Cancer. *Immunity* (2018) doi:10.1016/j.immuni.2018.03.023.
72. Yost, K. E. *et al.* Clonal replacement of tumor-specific T cells following PD-1 blockade. *Nat. Med.* (2019) doi:10.1038/s41591-019-0522-3.
73. Richardson, T. G., Harrison, S., Hemani, G. & Smith, G. D. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *Elife* (2019) doi:10.7554/eLife.43657.
74. Manoury, B. *et al.* An asparaginyl endopeptidase processes a microbial antigen for class II MHC presentation. *Nature* (1998) doi:10.1038/25379.
75. Alvarez-Fernandez, M. *et al.* Inhibition of mammalian legumain by some cystatins is due to a novel second reactive site. *J. Biol. Chem.* (1999) doi:10.1074/jbc.274.27.19195.
76. Huh, C. G. *et al.* Decreased metastatic spread in mice homozygous for a null allele of the cystatin C protease inhibitor gene. *J. Clin. Pathol. - Mol. Pathol.* (1999) doi:10.1136/mp.52.6.332.
77. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat. Protoc.* (2010) doi:10.1038/nprot.2010.116.
78. Sayaman, R. W. *et al.* Germline genetic contribution to the immune landscape of cancer. *Immunity* (2021) doi:10.1016/j.immuni.2021.01.011.
79. Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat. Methods* (2018) doi:10.1038/s41592-018-0051-x.
80. Loh, P. R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nature Genetics* (2018) doi:10.1038/s41588-018-0144-6.
81. Ghoussaini, M. *et al.* Open Targets Genetics: Systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Res.* (2021) doi:10.1093/nar/gkaa840.
82. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* (2015) doi:10.1038/ng.3404.
83. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput. Biol.* (2015) doi:10.1371/journal.pcbi.1004219.
84. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* (2016) doi:10.1038/ng.3538.
85. Nasser, J. *et al.* Genome-wide enhancer maps link risk variants to disease genes. *Nature* (2021) doi:10.1038/s41586-021-03446-x.
86. Gel, B. & Serra, E. KaryoploteR: An R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* (2017) doi:10.1093/bioinformatics/btx346.
87. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* (2009) doi:10.1093/bioinformatics/btp616.
88. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-Content Normalization for RNA-Seq Data. *BMC Bioinformatics* (2011) doi:10.1186/1471-2105-12-480.
89. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* (2021)

doi:10.1016/j.cell.2021.04.048.
90. Tickle T, Tirosh I, Georgescu C, Brown M, H. B. inferCNV of the Trinity CTAT Project.
91. Riemondy, K. A. *et al.* clustifyr: An R package for automated single-cell RNA sequencing cluster classification. *F1000Research* (2020) doi:10.12688/f1000research.22969.2.
92. Franzén, O., Gan, L. M. & Björkegren, J. L. M. PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database* (2019) doi:10.1093/database/baz046.

Author contributions

S.O.K., H.V.M. and T.J. conceived and designed the study. S.O.K. performed statistical and computational analyses. S.B. performed eQTL analyses in the STARNET cohort. H.L. contributed to data curation of clinical cohorts. B.D. performed *in vitro* experiments. M.F. performed *in vivo* experiments. T.H. contributed to computational analyses. A.R. and J.L.M.B. coordinated data access in STARNET cohort. B.W. provided input for interpretation of results and direction for further analyses. H.V.M and T.J. supervised the project and guided all data analyses. S.O.K., H.V.M. and T.J. wrote the manuscript with input from all authors. All co-authors approved the final version of the manuscript.

Data and code availability

Due to the data use agreements for the datasets analyzed in this manuscript, we are unable to directly share or distribute any patient-level data. To facilitate dataset requests from applicable data use committees, we provide all accession codes for all datasets relating to this manuscript in the Methods and in table S2. UK Biobank data can be requested through the application process detailed at https://www.ukbiobank.ac.uk/. Primary data for *in vitro* and *in vivo* experiments will be published on acceptance of the peer-reviewed manuscript. Jupyter Notebooks detailing the core genomic analyses performed (cohort quality control, GWAS, SEM, PRS derivation) will be made available on Github on acceptance of a peer-reviewed manuscript. Summary statistics for GWAS and PGS will be uploaded to GWAS Catalog and PGS Catalog respectively.

Acknowledgements

This work was conducted using the UK Biobank resource under application number 58510. We thank all patients and their families who have volunteered to participate in clinical research, without whom this study would not have been possible. T.J. acknowledges funding from Cold Spring Harbor Laboratory and grants from The Mark Foundation for Cancer Research (33300111) and Nancy Gay Foundation (58391101). HVM receives funding from the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory. S.K. is supported by the Starr Centennial Scholarship at the Cold Spring Harbor Laboratory School of Biological Sciences. M.F. is supported by the 'La Caixa' Foundation (ID 100010434) in the framework of the 'La Caixa' Fellowship Program under agreement LCF/BQ/AA18/11680037. S.B. is supported by the Medical Research Council and B.R.W. by the Wellcome Trust. Computational analyses were performed with assistance from the US National Institutes of Health Grant S10OD028632-01.

Baseline characteristics		
Age, years	Median	61
	Range	18-89
Gender, n (%)	Female	214 (31)
	Male	471 (69)
Tumor type, n (%)	Melanoma	373 (54)
	Urothelial	221 (32)
	Renal cell carcinoma (RCC)	60 (9)
	Non-small-cell lung cancer (NSCLC)	19 (3)
	Head and neck squamous cell carcinoma (HNSCC)	12 (2)
Treatment, n (%)	Anti-PD-1	320 (47)
	Anti-PD-L1	207 (30)
	Anti-CTLA-4	145 (21)
	Combined PD-1/PD-L1 and CTLA-4	13 (2)
Best overall response (RECIST), n (%)	Complete response (CR)	47 (7)
	Partial response (PR)	128 (19)
	Stable disease (SD)	125 (18)
	Progressive disease (PD)	301 (44)
	Unavailable	84 (12)
Durable clinical benefit, n (%)	Yes	285 (42)
	No	385 (56)
	Unavailable	15 (2)

Table 1. Summary of patient demographics in pan-immunotherapy (panIO) cohort. For details on the individual cancer immunotherapy trial cohorts making up this data see Table S2.

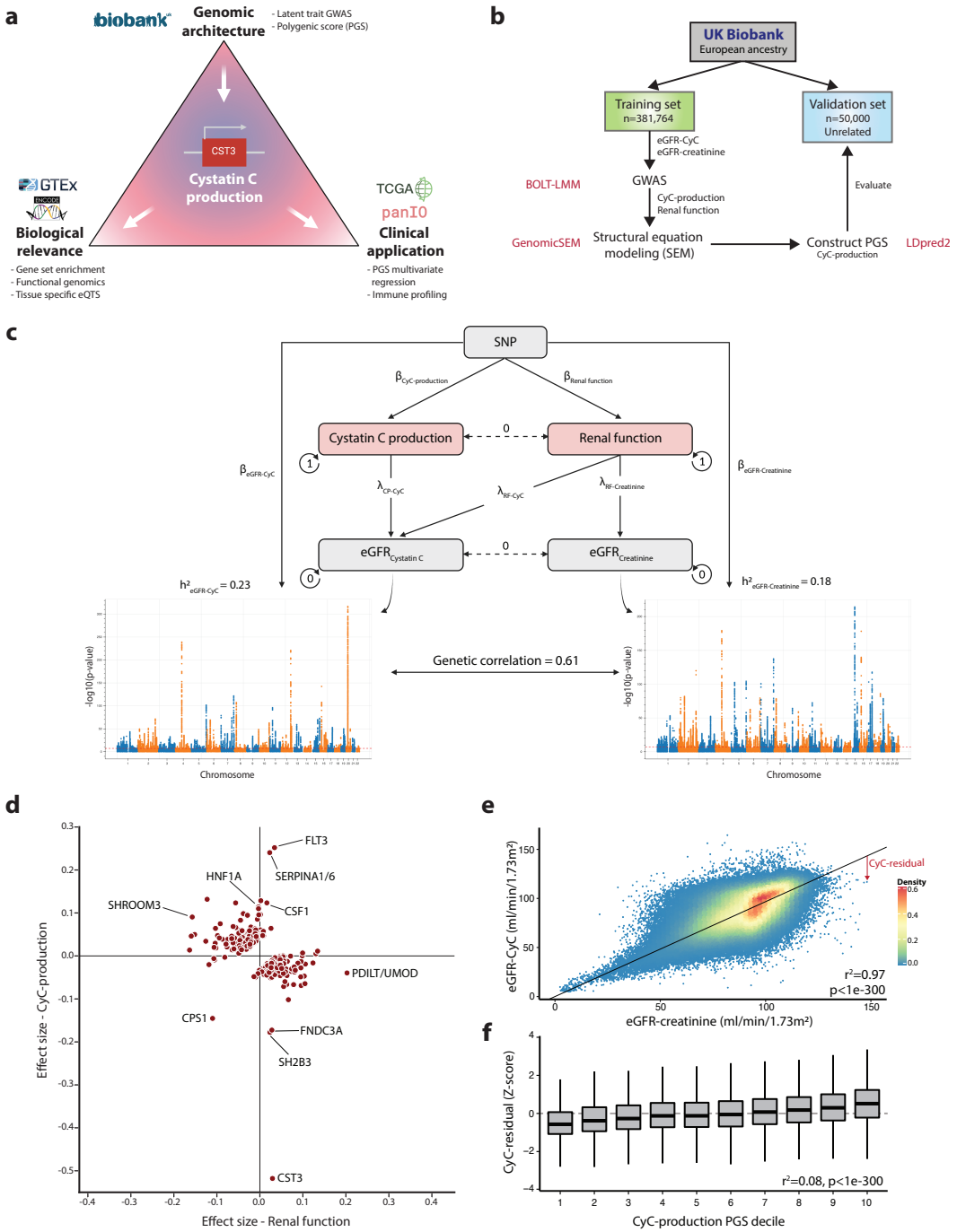

Figure 1. Genomic architecture of cystatin C (CyC) production. (a) Schematic of study plan. The analysis of CyC-production latent trait in UK Biobank (UKB) is leveraged to determine the biological and clinical relevance of CyC. (b) CONSORT diagram and summary of UKB genome-wide association analysis strategy in European population. The software packages utilized for each step are displayed in red. (c) Structural equation model to estimate latent traits of CyC-production and renal function. The model schematic, heritability (h^2) of eGFR-creatinine and eGFR-CyC, and their genetic correlation derived from LD score regression are shown. Circular arrows refer to variance of each component and dashed lines refer to covariance between components. RF, Renal Function. (d) Latent trait effect sizes (CyC-production and renal function) for index single nucleotide polymorphisms (SNPs). SNPs corresponding to each clumped locus in eGFR-CyC summary statistics are displayed. Gene names are annotated per OpenTargets V2G pipeline. (e) Linear model of eGFR-CyC as a function of eGFR-creatinine across all paired blood samples in UK Biobank, including sex as a covariate. The deviation of the eGFR-CyC from the linear fit as indicated by the red arrow is defined as the CyC-residual, a surrogate for CyC production. (f) Correlation of CyC-residual with CyC-production polygenic score (PGS). Only data from the independent validation set were used.

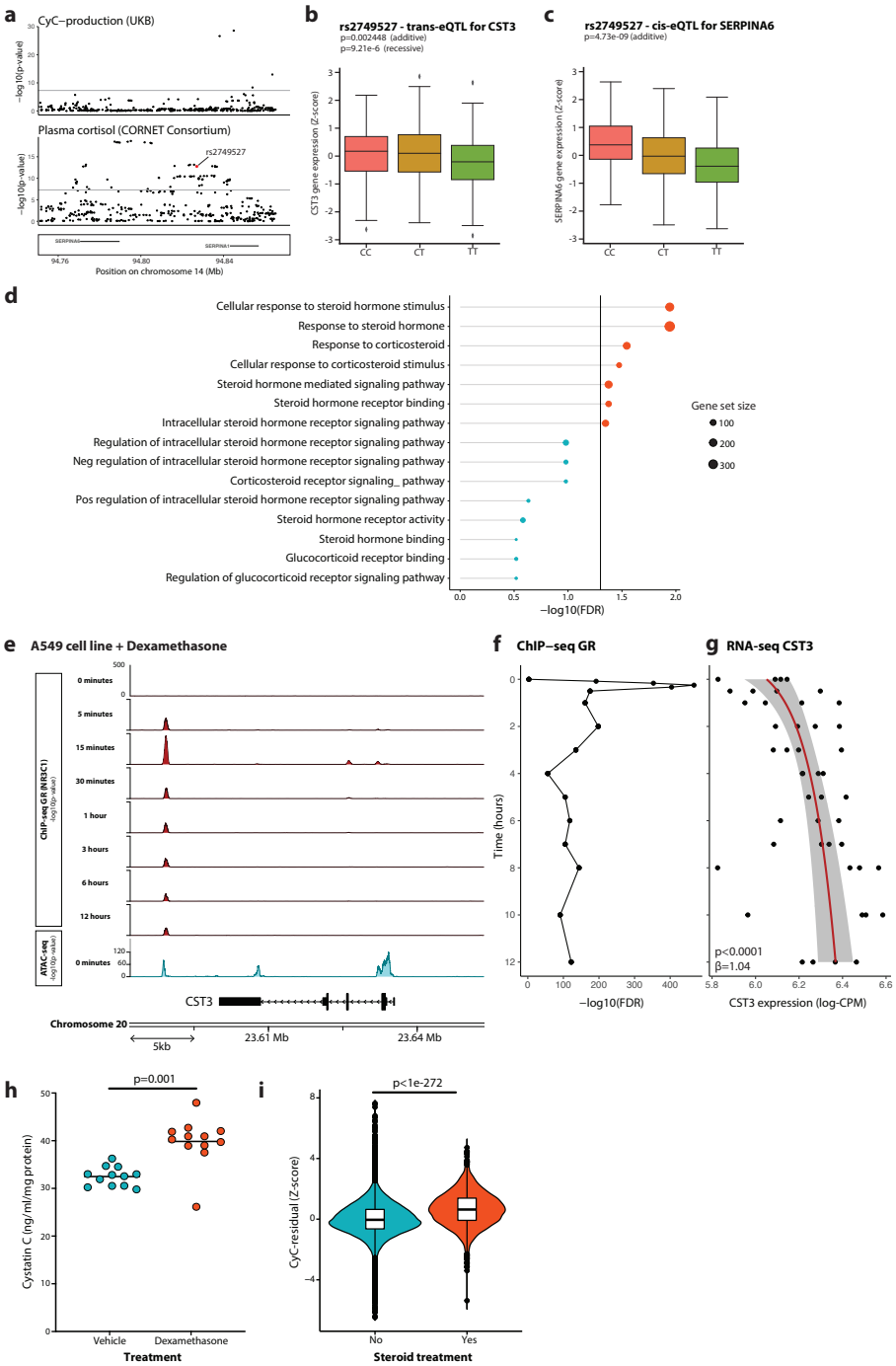

Figure 2. Cystatin C (CyC) is a glucocorticoid response gene. (a) Colocalization of summary statistics for CyC-production from UK Biobank (UKB) and plasma cortisol from CORNET Consortium at *SERPINA1/6* locus. rs2749527 variant is highlighted in red. (b) Trans-eQTL analysis examining association between genetic instrument rs2749527 and *CST3* gene expression in visceral adipose tissue (VAT) in STARNET cohort. (c) Cis-eQTL association between rs2749527 and *SERPINA6* (encodes cortisol-binding globulin) in liver in STARNET cohort. See Figure S3 for replication analysis in GTEX. (d) Gene set enrichment analysis (MAGMA) across CyC-production summary statistics for steroid signaling-related gene sets. Functional genomics in A549 cell line (ENCODE project) treated with 100nM dexamethasone for 0 minutes to 12 hours. (e) ChIP-seq (for glucocorticoid receptor/NR3C1) and ATAC-seq (at 0 hours) at *CST3* locus identifies a glucocorticoid-responsive and accessible distal enhancer element. Timecourse of (f) GR recruitment (at distal enhancer) and (g) *CST3* gene expression (log-CPM) following dexamethasone treatment in A549 cells (ENCODE project). Trendline and shaded 95% confidence interval correspond to regression of gene expression as a function of log-time. (h) Extracellular cystatin C concentration in A549 cells normalized to cellular protein content after 18-hour treatment with 100nM dexamethasone or vehicle control. Each condition comprises 12 biological replicates. See Figure S3 for timecourse. (i) Z-scored CyC-residual in UKB cohort stratified by oral steroid treatment.

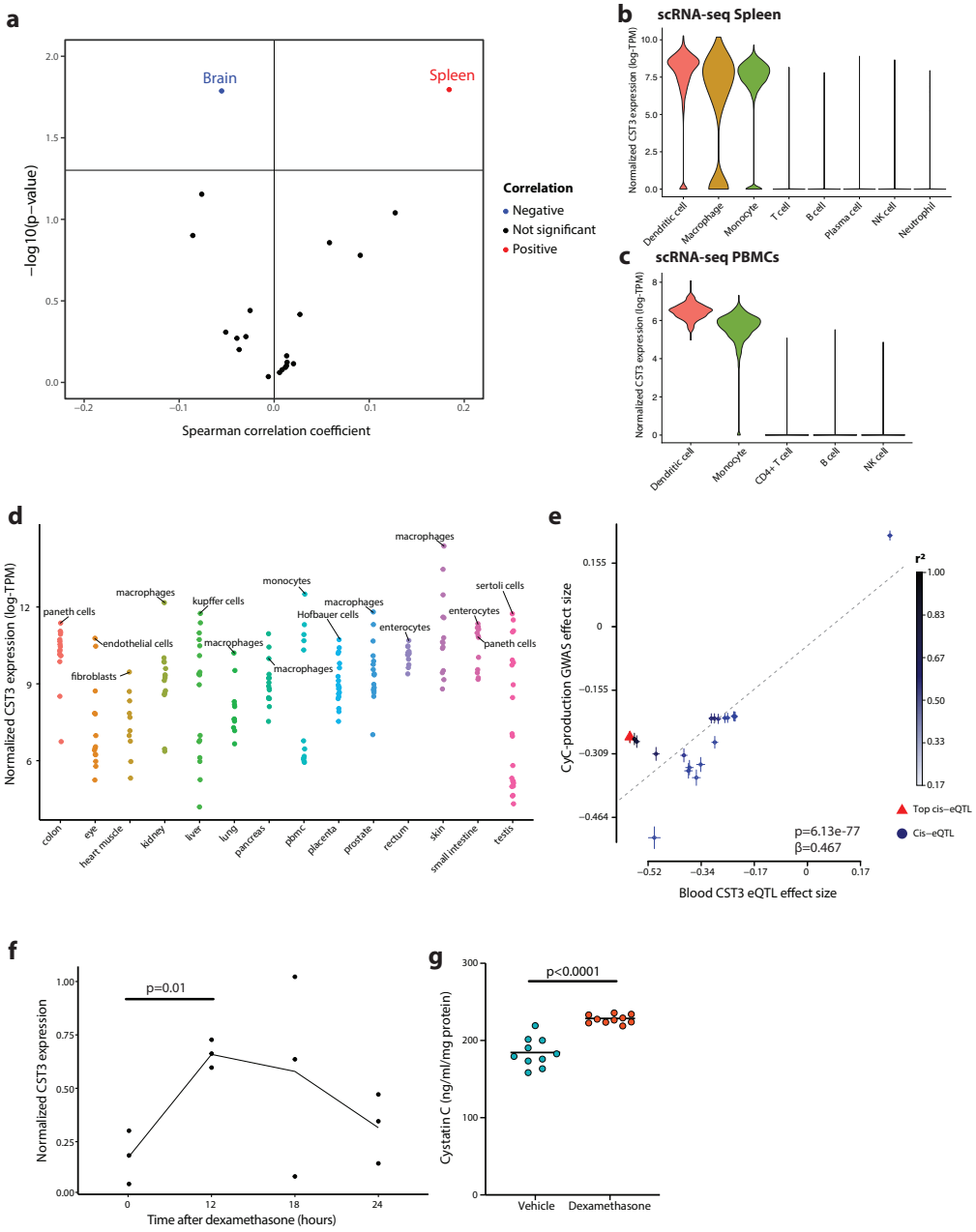

Figure 3. Cystatin C (CyC) is predominantly produced by myeloid cells in healthy individuals. (a) Tissue-specific expression qualitative trait score (eQTS) analysis to identify tissues with significant correlation (Spearman coefficient) between CyC-production polygenic score (PGS) and tissue-specific *CST3* gene expression in GTEX cohort. P values are uncorrected as each correlation test is performed in a non-overlapping set of tissue-specific samples. Distribution of normalized single-cell *CST3* expression (log-TPM) in cell clusters isolated from (b) spleen and (c) peripheral blood mononuclear cells (PBMCs). Clusters defined by correlation to reference PBMC data. (d) Mean *CST3* gene expression (log-TPM) in each cell cluster from multiple tissue-specific single-cell RNA sequencing projects, harmonized by Human Protein Atlas. The top cell cluster by tissue is annotated. (e) Two-sample mendelian randomization (MR) using blood-specific cis-eQTLs for *CST3* (eQTLGen) as exposure and CyC-production latent trait GWAS as outcome. Error bars correspond to standard errors, point color refers to linkage with top cis-eQTL. Change in (f) *CST3* gene expression (RT-PCR) and (g) extracellular CyC concentration in PMA-treated macrophage-like THP-1 cells normalized to cellular protein content after 18-hour treatment with dexamethasone (100nM) or vehicle control. Each condition comprises 10 biological replicates.

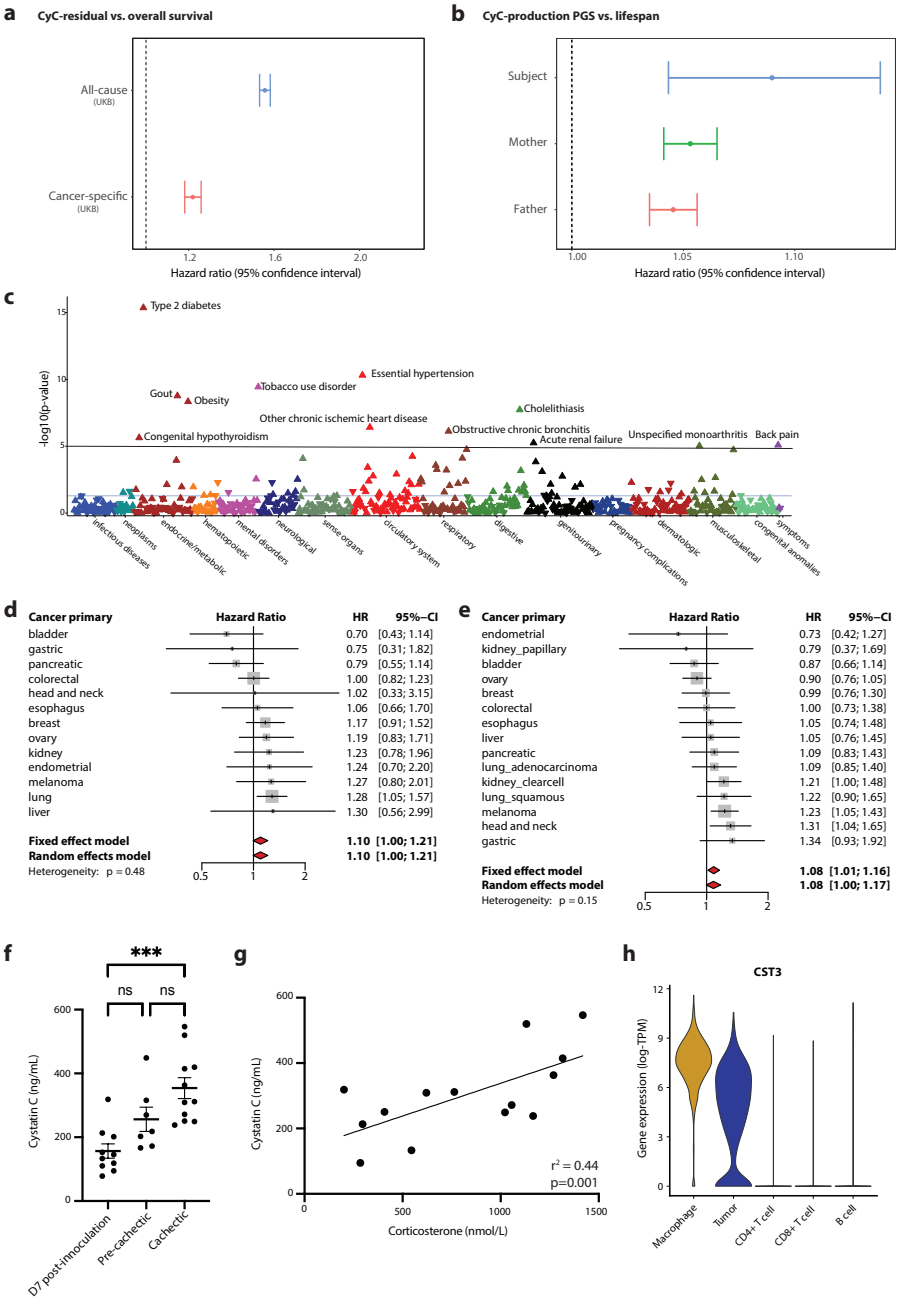

Figure 4. Cystatin C (CyC) is associated with multiple disease states and is prognostic in cancer patients. (a) Multivariate Cox regression to measure effect size for CyC-residual on overall survival in UK Biobank. Covariates included age, sex, body mass index (BMI), hemoglobin, C-reaction protein, eGFR-creatinine and operation status (for cancer-specific sub-analysis). (b) Multivariate Cox regression to measure effect size for CyC-production polygenic score (PGS) on subject and parental lifespan in UKB validation cohort. Covariates included PC1-4, recruitment center, genotyping array, year of birth of subject and sex of subject (if applicable). (c) Phenome-wide association (Cox regression) between CyC-production PGS and 694 time-to-event phenotypes in UKB validation cohort. Covariates included PC1-4, year of birth and sex (if applicable). Multivariate Cox regression to measure effect size for CyC-production PGS on disease-specific survival for specific cancers in (d) UKB validation cohort (cancers diagnosed since 2000, n=3954) and (e) TCGA cohort (n=4368). Covariates included PC1-4, age, sex (if applicable) and a term reflecting whether patient had curative surgery. (f) Plasma CyC concentration in BALBc mice after inoculation with colon-26 (C26) tumor cells. Cachexia defined by >15% body weight loss, pre-cachexia refers to 18 days after tumor inoculation. (g) Significant positive correlation between plasma corticosterone and plasma cystatin C during tumor progression in C26 model. (h) Single-cell CST3 gene expression in each cell cluster in melanoma tumors (n=12) from Jerby-Anon et al[70]. Clusters defined by correlation to reference PBMC data, with unclassified cells that exhibit detectable clonal copy number variation classified as tumor.

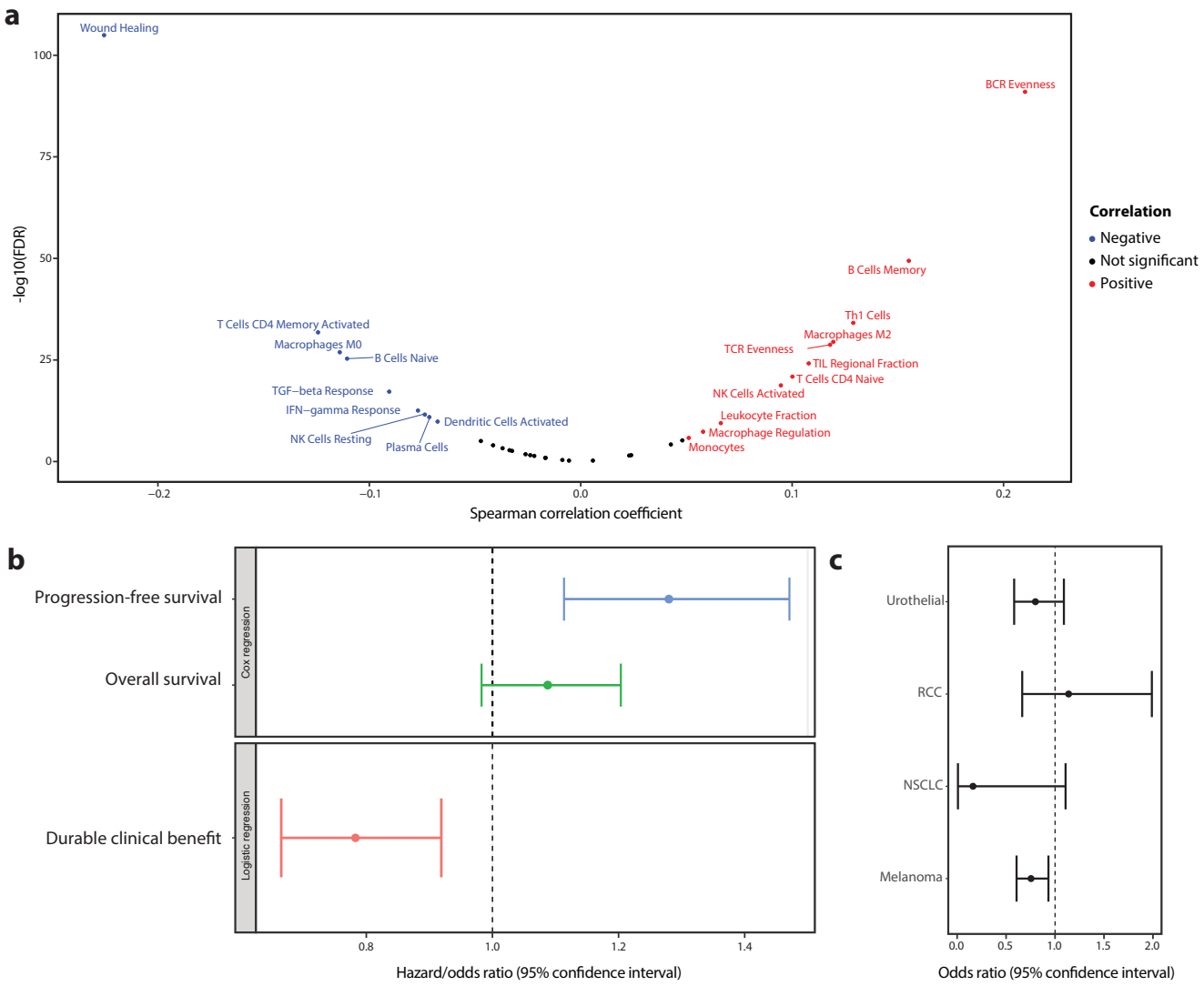

Figure 5. Cystatin C (CyC) is associated with failure of cancer immunotherapy. (a) Correlation (Spearman coefficient) between intratumoral CST3 expression and transcriptional immune signatures in TCGA cohort. High species evenness refers to reduced immune cell clonality. Significance defined by FDR < 0.05 and absolute coefficient > 0.05. (b) Multivariate (Cox and logistic) regression of Z-scored CyC-production polygenic score (PGS) against immuno-oncology outcomes (progression-free survival, overall survival, durable clinical benefit) in meta-analysis of European patients (n=685) treated with checkpoint immunotherapy (anti-CTLA4 or anti-PD1/PD-L1). Sample sizes for each clinical end-point were n=342, n=685 and n=670 respectively. In each model, covariates included the first four principal components, sex and cancer primary. Higher hazard ratios (survival) or lower odds ratios (durable clinical benefit) reflect worse therapeutic outcomes. (c) Sensitivity analysis indicating odds ratio and confidence interval for durable clinical benefit in each cancer type.

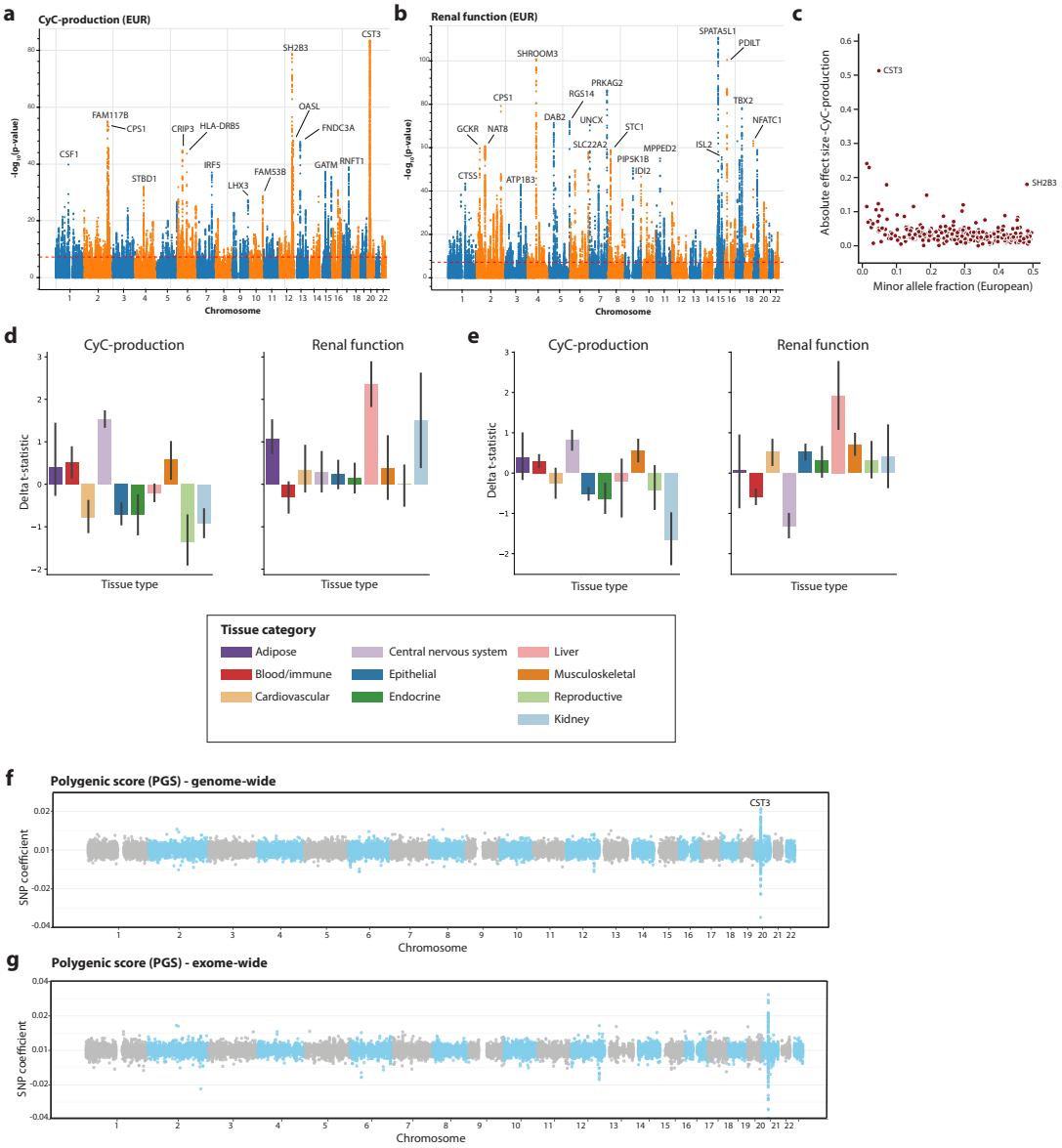

Figure S1. Summary statistics from (a) Cystatin C (CyC)-production and (b) Renal function latent traits in European UKB subjects, displayed a Manhattan plot. Loci with a p-value less than 1e-30 are annotated with gene name from OpenTargets V2G pipeline. (c) Relationship between effect size and minor allele frequency in CyC-production trait, annotated with outlier loci. Partitioned heritability analysis across multiple tissue types derived from (d) gene expression and (e) chromatin accessibility data. Delta t-statistic refers to change in enrichment t-statistic between measured eGFR-CyC summary statistics and latent CyC-production or renal function statistics. Errors bars signify 95% confidence interval. Coefficients for each SNP included in polygenic scores (PGS) generated using (f) HapMap SNPs (n=1,031,527) or (g) HapMap SNPs that can be reliably imputed from exome (n=352,549) sequencing data. CST3 locus on chromosome 20 is annotated.

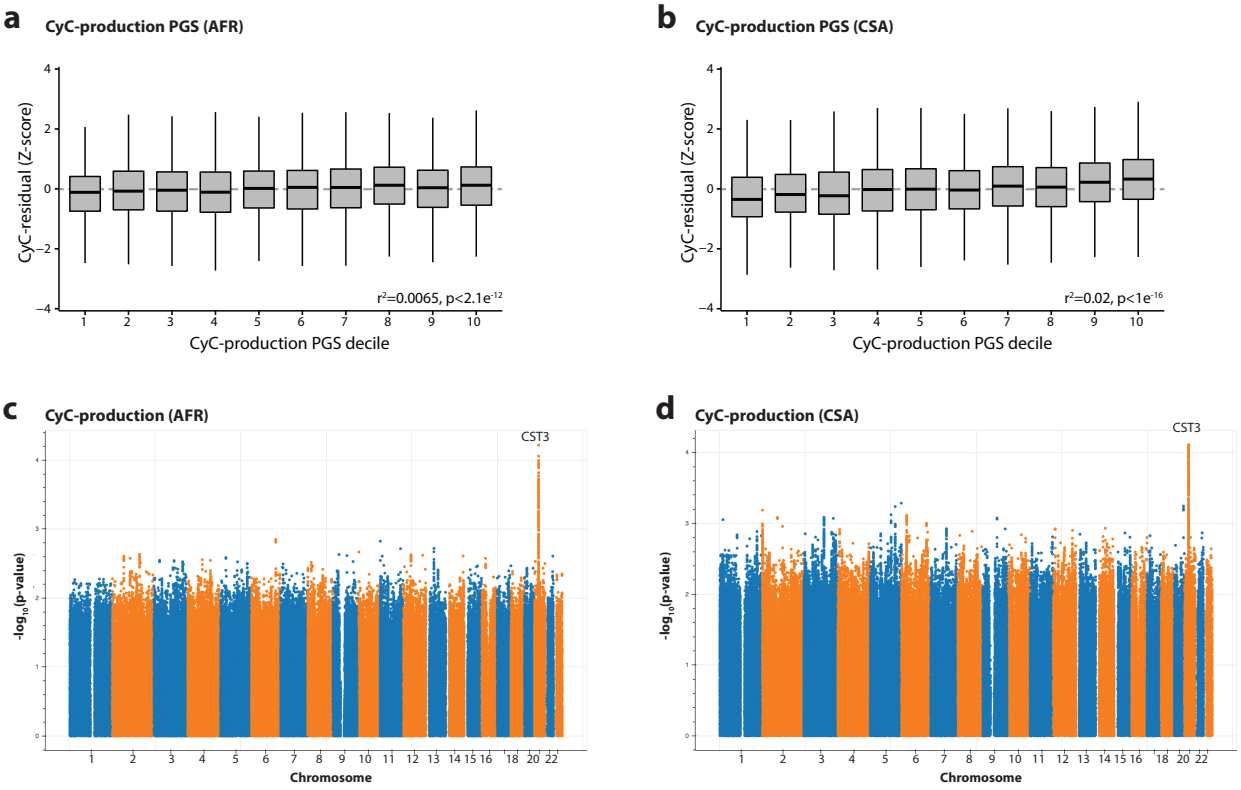

Figure S2. Trans-ancestral portability of CyC-production polygenic score (PGS) derived in UK Biobank European training set (n=381,764) applied to subjects of (a) African (AFR, n=8152) and (b) Central and South Asian (CSA, n=9845) genetic ancestry. Summary statistics from CyC-production latent trait in (c) AFR and (d) CSA genetic ancestry cohorts, derived from GWAS for eGFR-CyC and eGFR-creatinine followed by structural equation modeling. Results displayed as Manhattan plot; no loci reached genome-wide significance in latent trait analysis.

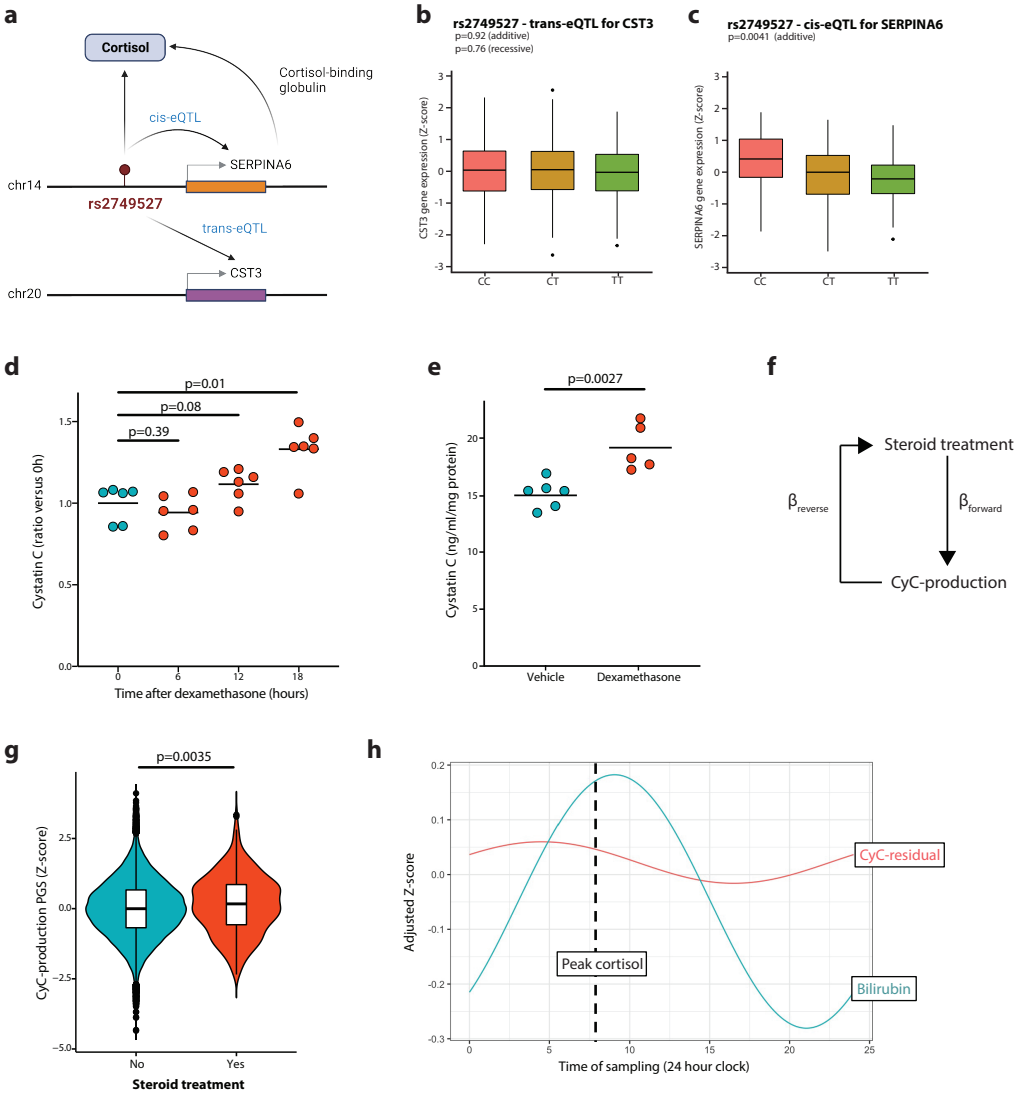

Figure S3. (a) Role of single genetic instrument, rs2749527, as a trans-eQTL for *CST3* on chromosome 20 (in visceral adipose tissue) and as a cis-eQTL for *SERPINA6* on chromosome 14 (in liver), which codes for cortisol-binding globulin, that is also significantly associated with morning plasma cortisol. Association analysis between rs2749527 and both (b) *CST3* gene expression in visceral adipose tissue (VAT, n=381) and (c) *SERPINA6* gene expression in liver (n=175) in GTEX cohort (EUR ancestry). (d) Ratio of extracellular cystatin C concentration in A549 cells normalized to cellular protein content after 0-, 6-, 12- or 18-hour treatment with dexamethasone. Each timepoint comprises 4 biological replicates. (e) Extracellular cystatin C concentration in HeLa cells normalized to cellular protein content after 18-hour treatment with dexamethasone or vehicle control. Each condition comprises 5-6 biological replicates. (f) Hypothetical model for the relationship between steroid treatment and CyC-production. In forward case, steroid treatment directly increases CyC-production. In reverse case, CyC-production germline predisposition influences probability of steroid treatment. (g) Distribution of Z-scored CyC-production polygenic score (PGS) in subjects in UK Biobank (UKB) validation set with/without a history of steroid treatment. (h) Diurnal variation in CyC-residual and bilirubin (known to correlate with plasma cortisol) derived from cosinor regression in UKB cohort. Adjusted Z-score refers to Z-scoring stratified by age and gender.

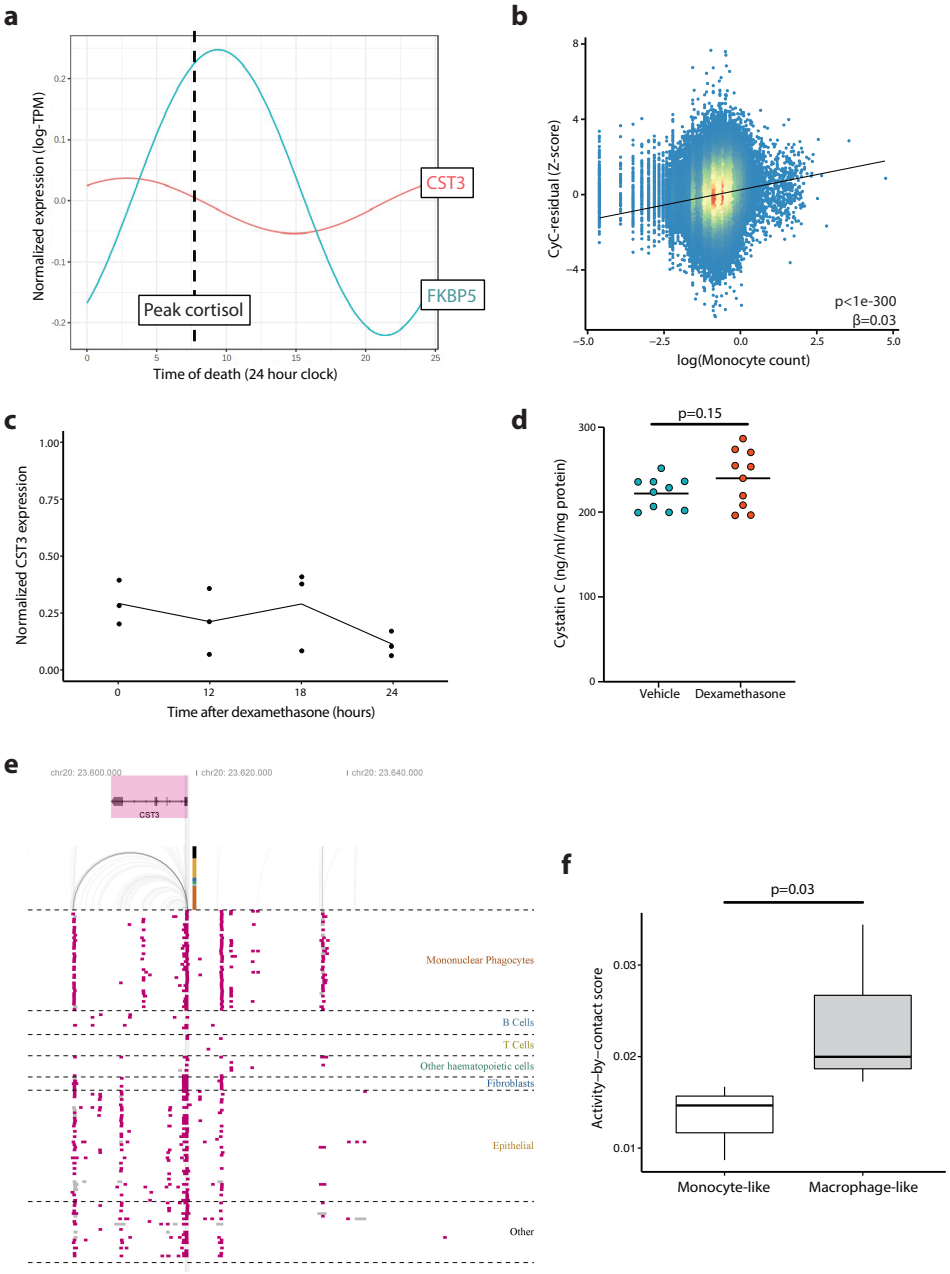

Figure S4. (a) Diurnal variation in *CST3* and *FKBP5* (canonical glucocorticoid response gene) derived from cosinor regression in GTEX spleen cohort, using time of death for each GTEX donor. (b) Significant positive correlation between logarithm of monocyte count and Z-score cystatin C (CyC)-residual in UKB cohort. Significance refers to multivariate regression including age, sex and body mass index (BMI). Change in (c) *CST3* gene expression (RT-PCR) and (d) extracellular CyC normalized to cellular protein content after treatment with 100nM dexamethasone or vehicle control in monocyte-like THP-1 cell line. Each condition comprises 10 biological replicates. (e) Visualization of predicted enhancer elements at CST3 locus from activity-by-contact (ABC) model[85], showing distal enhancer element acting on cystatin C. Each row corresponds to a cell line-treatment pair and epithelial grouping includes cancer cell lines. (f) ABC model scores for distal enhancer in THP-1 cells - with (macrophage-like) or without (monocyte-like) PMA treatment.

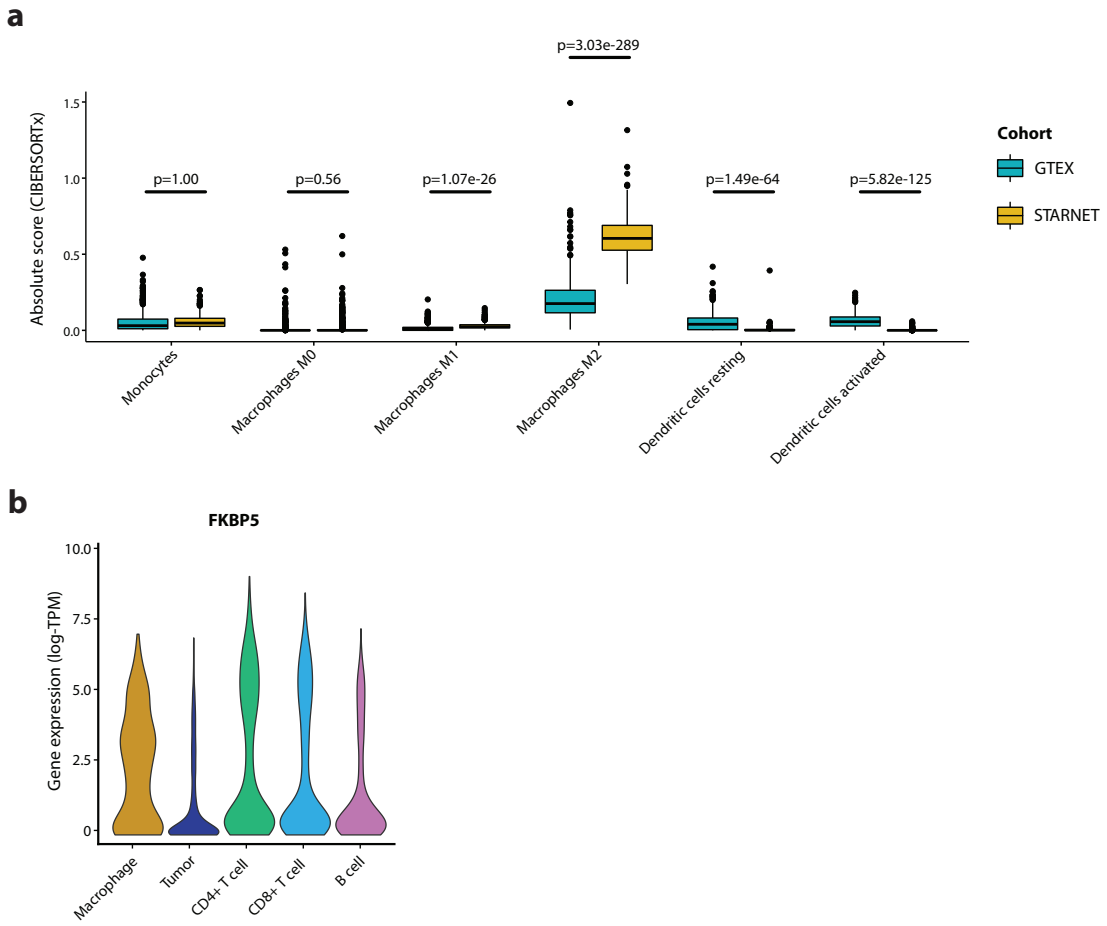

Figure S5. (a) Inferred absolute myeloid cell composition from CIBERSORTx analysis (absolute mode) applied to visceral adipose tissue (VAT) from STARNET and GTEX cohorts. p values refer to t-tests with Bonferroni correction. While units are comparable between cell types and samples, they do not refer to an absolute cell fraction. (b) Single-cell *FKBP5* (canonical glucocorticoid receptor target) gene expression in each cell cluster in melanoma tumors (n=12) from Jerby-Anon et al[70] Clusters defined by correlation to reference PBMC data, with unclassified cells that exhibit detectable clonal copy number variation classified as tumor.

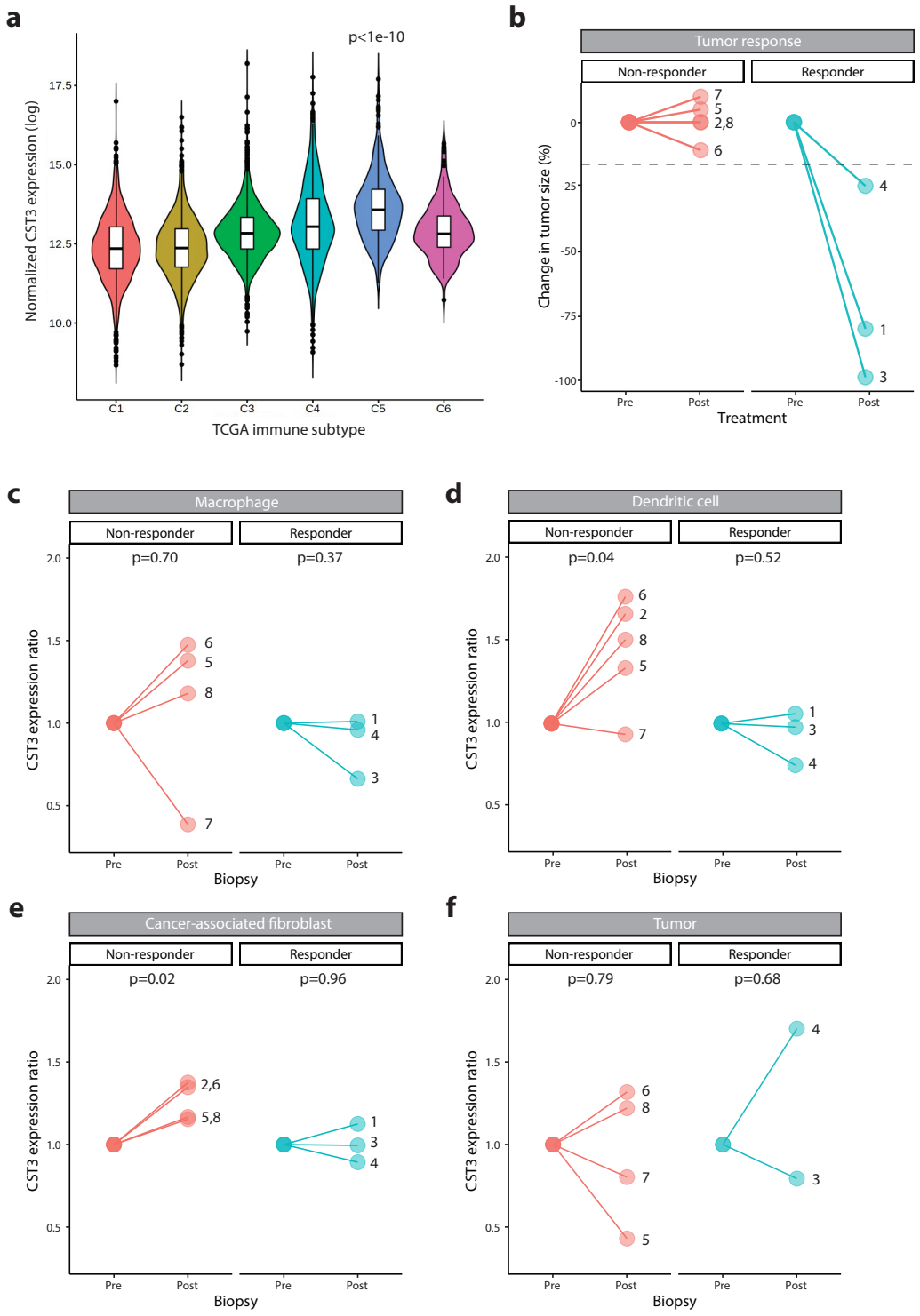

Figure S6. (a) *CST3* gene expression (log-normalized) in each TCGA immune subtype derived by Thorssen et al[70]. Subtype 5 corresponds to 'immunological quiet' subtype, characterized by reduced lymphocyte and increased M2 macrophage responses. Significance level refers to one-way ANOVA with post-hoc Tukey's test. (b) Summary of radiological response data for patients with basal cell carcinoma (BCC, n=8) derived from Yost et al[72]. Response is nominally defined as tumor regression >25%. Numbers refer to patient IDs. Ratio of cluster-specific pseudo-bulk *CST3* gene expression (log-TPM) in paired biopsy samples pre/post anti-PD-1 treatment, segregated according to clinical response, for (c) macrophage, (d) dendritic cell, (e) cancer-associated fibroblast (CAF) and (f) tumor subsets. Uncorrected P values refer paired t-tests.