

Urea cycle dysregulation generates clinically relevant genomic and biochemical signatures

Joo Sang Lee^{1,2*}, Lital Adler^{3*}, Hiren Karathia², Narin Carmel³, Shiran Rabinovich³, Noam Auslander^{1,2}, Rom Keshet³, Noa Stettner^{3,4}, Alon Silberman³, Lilach Agemy⁵, Daniel Helbling⁶, Raya Eilam⁴, Qin Sun⁷, Alexander Brandis⁸, Sergey Malitsky⁸, Maxim Itkin⁸, Hila Weiss³, Sivan Pinto³, Shelly Kalaora⁹, Ronen Levy⁹, Eilon Barnea¹⁰, Arie Admon¹⁰, David Dimmock¹¹, Noam Stern-Ginossar¹², Avigdor Scherz⁴, Sandesh CS Nagamani^{7,13}, Miguel Unda^{14,15}, David M Wilson III¹⁶, Ronit Elhasid¹⁷, Arkaitz Carracedo^{15,18,19,20}, Yardena Samuels⁹, Sridhar Hannenhalli², Eytan Ruppin^{1,2,21**} and Ayelet Erez^{3**}

¹Cancer Data Science Lab, National Cancer Institute, National Institute of Health, Bethesda, MD 20892, USA;

²Center for Bioinformatics and Computational Biology, University of Maryland Institute for Advanced Computer Studies, Dept. of Computer Science, University of Maryland, College Park, MD, 20742, USA;

³Department of Biological Regulation, Weizmann Institute of Science, Rehovot, 7610001, Israel;

⁴Department of Veterinary Resources, Weizmann Institute of Science, Rehovot, 7610001, Israel;

⁵Department of Plant and Environmental Science, Weizmann Institute of Science, Rehovot, 7610001, Israel;

⁶Envision Genomics, Huntsville, Alabama, 35806, USA;

⁷Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, 77030, Texas; USA;

⁸Life Sciences Core Facilities, Weizmann Institute of Science, Rehovot, 7610001, Israel;

⁹Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, 7610001, Israel;

¹⁰Faculty of Biology, Technion – Israel Institute of Technology, Haifa, 3200003, Israel;

¹¹Rady Children's Institute for Genomic Medicine, San Diego, CA 92123 USA;

¹²Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, 7610001, Israel;

¹³Texas Children's Hospital, Houston, 77030, Texas; USA;

¹⁴Department of Urology, Basurto University Hospital, 48013, Bilbao, Spain;

¹⁵CIBERONC;

¹⁶Laboratory of Molecular Gerontology, National Institute on Aging, Intramural Research Program, National Institutes of Health, 251 Bayview Blvd., Baltimore, MD 21224

¹⁷Pediatric Hemato Oncology Department, Sourasky Medical Center, affiliated to the Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv, 6997801, Israel;

¹⁸CIC bioGUNE, Bizkaia Technology Park, 801 building, 48160, Derio, Spain;

¹⁹Ikerbasque, Basque Foundation for Science, Bilbao, Spain;

²⁰Biochemistry and Molecular Biology Department, University of the Basque Country (UPV/EHU), Bilbao, Spain;

²¹Schools of Medicine and Computer Science, Tel-Aviv University, 6997801, Israel.

*, **These authors contributed equally to this work

**Correspondence to/ Lead contacts: Ayelet Erez MD, PhD, Weizmann Institute of Science, Rehovot, Israel; Phone- 972-8-934-3714; FAX- 972-8-934-3739; email: ayelet.erez@weizmann.ac.il &

Eytan Ruppin MD, PhD, National Cancer Institute, National Institute of Health, Bethesda MD; Phone-1-240-858-3169; FAX-1-301-496-2905; email:eyruppin@gmail.com

Summary

The urea cycle (UC) is the main pathway by which mammals dispose waste nitrogen. We find that specific alterations in the expression of most UC enzymes occurs in many tumors, leading to a general metabolic hallmark termed UC dysregulation (UCD). UCD elicits nitrogen diversion towards carbamoyl-phosphate synthetase2, aspartate transcarbamylase and dihydroorotase (CAD) activation and enhances pyrimidine synthesis, resulting in detectable changes in nitrogen metabolites in both patient tumors and bio-fluids. The accompanying excess of pyrimidine *vs* purine nucleotides results in a genomic signature consisting of transversion mutations at the DNA, RNA and protein levels. This mutational bias is associated with more hydrophobic tumor antigens and with a better response to immune checkpoint inhibitors, independent of mutational load. Taken together, our findings demonstrate that UCD is a common feature of tumors, which profoundly impacts carcinogenesis, mutagenesis and immunotherapy response.

Introduction

In the liver, the urea cycle (UC) converts excess systemic nitrogen, derived from the breakdown of nitrogen-containing metabolites as ammonia and glutamine to urea, a disposable nitrogenous compound (Ah Mew et al., 1993). Outside the liver, different UC enzymes are expressed in accordance with cellular needs for UC intermediates. Mendelian disorders with deficiencies of UC enzymes and transporters were recognized many years ago (Ah Mew et al., 1993). Whereas these UC inherited disorders were not associated with cancer predisposition, anecdotal studies have reported the altered expression of specific UC components in cancer (Chaerkady et al., 2008; Lee et al., 2014; Syed et al., 2013). We have shown that the loss of UC enzyme argininosuccinate synthase (ASS1) promotes cancer proliferation by diversion of its aspartate substrate towards carbamoyl-phosphate synthase 2 (CPS2), aspartate transcarbamylase (ATC), and dihydroorotase, the CAD enzyme that catalyzes the first three reactions in the pyrimidine synthesis pathway (Nagamani and Erez, 2016; Rabinovich et al., 2015). Similarly, it was shown that the UC enzyme CPS1 maintains the pyrimidine pool in non-small cell lung cancer through CAD activation (Kim et al., 2017). Based on these recent discoveries of UC rewiring towards pyrimidine synthesis and the dependence of tumors on UC nitrogen sources (Spinelli et al., 2017; Wise and Thompson, 2010), we hypothesized that UC dysregulation (UCD) maybe a widespread advantageous metabolic phenomenon in cancer (**Figure 1A**). As such, unravelling the molecular consequences of UCD in cancer may hold promising diagnostic and therapeutic opportunities.

Results

UCD induces CAD activation and facilitates cell proliferation

UC enzymes compete for their nitrogenous substrates with other enzymes such as CAD. Previous work on ASS1 and CPS1 (Kim et al., 2017; Rabinovich et al., 2015) demonstrated that alterations in specific UC enzymes' expression would increase nitrogenous substrate availability for CAD and pyrimidine synthesis. Such a consequence could predictably be also achieved by downregulation of the expression of argininosuccinate lyase (ASL), ornithine transcarbamylase (OTC), and ORNT1 (*SLC25A15*, the transporter which imports ornithine to the mitochondria), or from the overexpression of the UC transporter citrin (*SLC25A13*) that exports aspartate from the mitochondria (**Figure 1A and Figure S1A**). We first explored this hypothesis in patients with UC-

related inborn errors of metabolism, to understand whether monogenic alterations in UC enzymes would result in the anticipated activation of CAD. Indeed patients with germline mutations in OTC, ASL, ASS1, or ORNT1 have reportedly increased pyrimidine-related metabolites in their plasma or urine, whereas patients with CPS1 mutations do not (Brosnan and Brosnan, 2007; Gerrits et al., 1993; Salerno et al., 1999; van de Logt et al., 2017; van Kuilenburg et al., 2006). Corroboratively, we find that fibroblasts from patients with OTC and ORNT1 deficiencies exhibit higher proliferation rates and increased CAD activation compared to non-mutated controls, as exemplified by higher CAD activating phosphorylation (Ben-Sahra et al., 2013), while CPS1 deficient fibroblasts did not (**Figure S1B**). In another non-cancerous setting, we find that infecting non-transformed cells with Cytomegalovirus (CMV) decreases *ASS1* expression and elevates of *SLC25A13* levels (**Figure S1C**). This may explain the resultant documented activation of CAD and expansion of pyrimidine pools following viral infection (DeVito et al., 2014). Additionally, we find a significant increase in proliferation following induction of UCD by either downregulating *ASS1* or upregulating citrin in MCF-10 non-transformed epithelial cells (**Figure S1D**). Finally, a genome scale metabolic modelling analysis further supports the notion that UCD would lead to diversion of nitrogenous metabolites towards the synthesis of pyrimidines (**Figure S1E-F**).

To test the relevance of our findings to cancer, we screened different tumor cell lines and chose those in which the basal expression of UC enzymes enabled us to elicit and study UCD; we downregulated OTC in hepatocellular carcinoma (Hepg2), ORNT1 (*SLC25A15*) in ovarian carcinoma (SKOV), and overexpressed citrin in melanoma cells (LOX IMVI) (**Figure S2A**). Following each specific perturbation, we measured CAD activation through its phosphorylation on serine 1859 (Ben-Sahra et al., 2013). Each of these separate perturbations led to an increase in CAD phosphorylation and enhanced cellular proliferation (**Figure S2A**). CAD activation by phosphorylation is mediated by pS6K, following mammalian target of rapamycin (mTOR) pathway activation (Ben-Sahra et al., 2013; Robitaille et al., 2013). Indeed we find that UCD induces pyrimidine synthesis in an mTOR dependent manner, as evident from the increase in the levels of pS6K and pCAD (**Figure S2A**). Reintroducing the baseline expression of the perturbed UC enzymes to the cancer cells, as well as treating the cells with the mTOR inhibitor Rapamycin, reduced CAD phosphorylation and restricted proliferation (**Figure S2B-C**). Additionally,

inhibiting thymidylate synthase *via* Fluorouracil (5FU) significantly decreased proliferation and restricted the enhanced pyrimidine synthesis in the UC perturbed cells, showing that the effects of these UC perturbations is indeed mediated *via* increased pyrimidine synthesis (**Figure S2D**). The downregulation of OTC and ORNT1 resulted in increased ¹⁵N labelling of uracil from glutamine *in vitro* and increased tumor growth *in vivo* (**Figure S2E**), akin to the findings of (Rabinovich et al., 2015). In addition, in liver cancer cells, the knockdown of OTC significantly increased the uracil to urea ratio, reflecting a shift of nitrogen for pyrimidine synthesis over its catabolism to urea (**Figure S2F**). These findings markedly extend upon previously published results (Rabinovich et al., 2015) (Kim et al., 2017), showing that the specific perturbations of UCD enzymes results in CAD activation and increased proliferation in both non-cancerous and cancerous cells.

UCD is prevalent in cancer and associates with worse patient prognosis

UCD and the consequent flux of nitrogen towards CAD can thus be achieved through specific alterations in the expression of different enzymes in the cycle (**Figure S1A**). We hence designed a *UCD score* that captures the magnitude of the expression changes in UC components that are likely to lead to CAD activation (see **Methods**). To study how prevalent is UCD in tumor samples, we analyzed transcriptomics data to compute UCD scores in tumor specimens from the cancer genome atlas (TCGA) and in the corresponding healthy tissue samples (obtained from the Genotype-Tissue Expression project (GTEx) (Consortium, 2013,)) (**Supplemental Data**). We find that most tumor types (23 out of 25) have aberrant expression of at least two UC components in the direction expected to facilitate the supply of metabolites for CAD (**Figure 1B, Table S1**). Moreover, high UCD scores are associated with higher CAD expression (which has also been associated with CAD activation (Richmond et al., 2012)), higher cancer cell proliferation markers (by means of Proliferating Cell Nuclear Antigen, PCNA abundance and Ki-67 levels), increased tumor de-differentiation and grading and, importantly, with reduced patient survival (**Figure 2A-E, Figure S3**).

UCD alters nitrogen composition in mouse cancer models and in cancer patients' bio-fluids.

To further study the robustness of UCD in cancer patients, we assessed whether changes in pyrimidine synthesis in tumors could be detected in urine of cancer patients and found significantly elevated pyrimidines in urine of prostate cancer patients compared to healthy

controls (**Figure 3A**). We additionally find significantly higher pyrimidine to purine metabolites ratios in tumors with high vs. low UCD scores in human hepatocellular carcinoma (Budhu et al., 2013) and breast cancer (Terunuma et al., 2014) (**Figure 3B**). These findings are recapitulated in the urine of mice bearing breast cancer as well (**Figure S4A**). Inducing colon cancer in mice, we find that UCD occurs endogenously and is associated with increased pyrimidine synthesis-related metabolites in the urine of tumor bearing mice, compared to healthy controls (**Figure 3C and Figure S4B-C**). Although we found increased amino acids secretion in urine of mice bearing tumors relative to controls, amino acid secretion in the urine of mice with UCD-tumors was lower compared to mice bearing non-UCD-tumors (**Figure S4D-E**). As another readout for nitrogen homeostasis following tumorigenesis, we measured plasma urea levels in mice and patients bearing tumors. We found that plasma urea levels are lower in mice bearing tumors compared to control mice (**Figure 3D**), and observed a decreased expression of UC enzymes in the liver in mice bearing tumors (**Figure S4F**), suggesting a shift from catabolizing nitrogen to urea, to funneling it to tumor pyrimidine synthesis. To evaluate the relevance of our findings in humans, we surveyed the medical records of 100 pediatric cancer patients in a large medical center in Israel for their plasma urea levels on admission day, and compared them to age and gender matched mean urea values of 1.5 million Israeli healthy children (to overcome the age-related increase in plasma urea levels (Musch et al., 2006). Across a broad array of cancer types, we found that children with cancer have significantly decreased plasma urea levels compared with age-matched healthy controls (**Figure 3E**). Taken together, these findings testify that tumor UCD favours pyrimidine synthesis and is associated with identifiable alterations in nitrogen metabolites in mice and cancer patients' bio-fluids.

UCD induces a pyrimidine-rich transversion mutational bias

To study the effect of UCD on cancer genome, we further analyzed the UCD-induced cancer cells that we generated by perturbing UC enzymes. First, as expected, we find an increase in the ratio of pyrimidine to purine metabolites (**Figure 4A, Figure S4E and Table S2**). Second, since the nucleotide pool imbalance has been associated with mutagenesis (Schmidt et al., 2017), we sequenced the genome of these UCD cellular cancers to identify the genomic ramifications of UCD (**Table S3**). We detected a significant bias toward R->Y (puRine to pYrimidine) vs Y->R point mutations on the DNA sense strand, which we termed a Pyrimidine-

rich Transversion Mutational Bias (PTMB) (**Figure 4B, Methods**). This mutational bias continues to accumulate and increase over time in UCD-induced cells, as evident by comparing the PTMB levels of these cells after 1 and 2 weeks post UCD induction (see **Methods, Figure 4C**, also in mouse data, **Figure S5A**) and is independent of mutation load (see **Methods**, also in patient data, **Figure S5B**).

Analyzing the TCGA data, we find that PTMB is quite prevalent in patients' tumors (**Table S1**), and is significantly associated with UCD across different cancer types (**Figure 5A-B, Figure S5C-E, Supplemental Data**). PTMB is not strongly associated with known mutational signatures (Alexandrov et al., 2013) including the well-established smoking signature (Alexandrov et al., 2016) that is enriched with transversion mutations (see **Methods** for details, **Figure S5D-E**). Notably, we find a highly significant correlation between gene expression levels and PTMB levels in UCD samples (**Figure 5C, Figure S5F-G**). This association, combined with the occurrence of PTMB on the sense strand, suggests that PTMB may emerge due to DNA damage associated with transcription (Jinks-Robertson and Bhagwat, 2014). The damaged DNA is then processed incorrectly due to error-prone DNA repair events stemming from the nucleotide pool imbalance (Berquist and Wilson, 2012). This eventually may lead to mutagenesis within actively transcribed genes on the sense strand, as previously reported (Haradhvala et al., 2016) (see **Methods** for a more detailed discussion of these hypothesized mechanisms). Interestingly, we observe that the association between UCD and PTMB is more significant for nonsynonymous mutations, further testifying to its functional importance (**Figure S5H**). Importantly, PTMB in mutated genes is also detectable in their corresponding mRNA and protein sequences, as we find analyzing pertaining multi-omics breast cancer data (Cancer Genome Atlas, 2012; Mertins et al., 2016) (**Figure 5D-E, Supplemental Data**). Together, these results demonstrate that UCD induces a unique PTMB signature in cancer that propagates from the DNA to mRNA to protein levels. PTMB is associated with worse patients' survival (**Figure 5F, Table S4**) and is distinct from other published mutational signatures (Alexandrov et al., 2016; Alexandrov et al., 2013).

UCD and PTMB are associated with better response to immune checkpoint therapies

Immune checkpoint therapy (ICT), an emerging anti-cancer treatment strategy, has proven beneficial in patients with tumors exhibiting high mutation load (Eggermont et al., 2016; Rizvi et al., 2015; Rosenberg et al., 2016; Snyder et al., 2014). Yet, tumors containing equally high mutation loads exhibit a variable response, suggesting that additional factors contribute to the overall immunotherapy success (Gubin et al., 2015; McGranahan and Swanton, 2017; Van Allen et al., 2015) (**Methods**). UCD-elicited PTMB led us to study the effects of UCD and PTMB on the efficacy of ICT, to understand whether UCD could be one such factor. We analyzed the transcriptomics of melanoma tumors from patients treated with ICT to compute the tumor's UCD scores (Van Allen et al., 2015); (Hugo et al., 2016; Roh et al., 2017). Interestingly, responders to anti-PD1 (Hugo et al., 2016) and anti-CTLA4 (Van Allen et al., 2015) therapy, harboured significantly higher tumor UCD-scores than non-responders (**Figure 6A, Figure S6A**), (this separation is higher than that obtained using the CAD expression levels themselves (**Figure S6B**).) To study the effects of PTMB levels on ICT response, we analyzed a large exome sequencing cohort of melanoma patients treated with anti-PD1 (Roh et al., 2017). Notably, we find that PTMB is a stronger predictor of response to anti-PD1 therapy than mutational load in these patients (**Figure 6B, Table S5**).

Tumor-specific mutations can result in immunogenic neo-antigens, which have also been correlated with responsiveness to ICT in highly mutagenic cancers (Miller et al., 2017). To learn more about the potential mechanisms underlying the increased ICT response associated with UCD and PTMB in patients, we performed an HLA peptidomics analysis of the UCD/ high PTMB cancer cells we have generated (as in **Figure 4B**) vs their WT parallels. We find that the UCD perturbed cancer cells have a larger number of R->Y than Y->R candidate neopeptides (15 vs 1 in total, and 4 vs 0 among those that are predicted to bind to MHC-I). The predicted MHC-I bound R->Y neopeptides are more hydrophobic (paired one-sided t-test $P < 0.04$, **Figure 6C, Table S6-7 and Methods**), and are hence expected to incur stronger immunogenicity (Chowell et al., 2015). Indeed, R->Y mutations are more likely to generate hydrophobic amino acids than other types of point mutations (Fisher exact test $P < 9.5E-5$, odds ratio=2.67, **Table S8 and Methods**). Interestingly, we also observe that UC dysregulated cells display more non-mutated hydrophobic peptides than control cells (**Figure S6C-D, Supplemental Data**). Thus, the

immunogenicity of the UC dysregulated cells may stem from the increased hydrophobicity of both their overall peptide repertoire and the PTMB-derived neo-antigens.

Given the strong predictive signal of UCD and PTMB levels on patients' response to ICT treatments reported above, we induced UCD in a syngeneic mouse model of colon cancer by knocking down ASS1 (**Figure S6E-F**). As previously reported by us (Rabinovich et al., 2015), we observed that this UC perturbation results in larger tumors *in vivo* (**Figure S6G**), as expected given the increased proliferation observed in UCD induced cancer cell-lines. We further find that ASS1 perturbed tumors are significantly more sensitive to anti-PD1-based ICT than the unperturbed ones (**Figure 6D**). As found in other studies (Wei et al., 2017), this increased therapeutic response is associated with enhanced specific infiltration of CD8 cytotoxic T-cells and not CD4 helper-T cells (**Figure 6E, Figure S6H**). Notably, the response to anti-PD1 treatment observed in mice bearing ASS1 knockdown tumors is characterized by a significantly attenuated cancer progression compared to that observed in the mice bearing unperturbed control tumors (**Figure 6F, Figure S6I**).

Discussion

Our results demonstrate that UCD exists across different cancer types, altering nitrogen utilization towards pyrimidine synthesis and supporting cancer proliferation. The resulting increase in cellular pyrimidines is associated with elevated R->Y transversions at the DNA levels, eliciting a transversion bias (PTMB) on the sense strand that is associated with worse prognosis. PTMB propagates from the DNA to the RNA and protein levels, leading to the generation of peptides with increased predicted immunogenicity. Inducing UCD enhances the response to ICT independently of mutational load in mouse models; analogously, increased UCD levels are associated with enhanced ICT response in patients. While nitrogen availability is an established promotor of proliferation, we show here that metabolic regulation of its fate by UC enzymes contributes to another cancer hall mark, which is mutagenesis. Encouragingly, mutagenesis which promotes carcinogenesis, also increases the beneficial response to immunotherapy. Based on our results, one may hypothesize that drugs that specifically promote nucleotide imbalance such as methotrexate (Allegra et al., 1987), could effectively increase the production of immunogenic neo-antigens and enhance the efficacy of ICT, especially in high UCD tumors. Furthermore, inducing

UCD in tumor by itself might be beneficial in increasing the response to ICT. Notably, in addition to the metabolic reprogramming of nitrogen in the tumors, we find that UCD is associated with other, systemic alterations in nitrogen homeostasis. Those include changes in the expression of UC genes in liver of tumor bearing mice and with decreased plasma urea levels in mice with cancer and in cancer patients at diagnosis. These findings require further explorations in larger animal and patient cohorts. Yet, they might indicate that cancers possibly actively modulate the host metabolism to increase their nitrogen supply and promote their growth and progression (**Figure 7**).

In sum, our results point to the important role of UCD in a broad spectrum of cancers, to the potential use of UCD-related metabolites as cancer biomarkers for early detection and lastly, to the role of UCD in predicting and potentially increasing the response to ICT.

Acknowledgments:

We acknowledge and thank the Weizmann Institute for providing financial and infrastructural support. We are grateful to the Clalit-Weizmann joint analysis program (G. Barbash, R. Balicer, A. Tanay), and in particular to Netta Mendelson Cohen, for providing access to distributions of Urea levels in healthy individuals. We thank Igor Ulitsky, Ron Rotkopf, Shifra Ben-Dor, Erez Persi, Edward M. Gertz, Sivan Galai for the intellectual and technical assistance. We are thankful to the Basque Biobank for Research (BIOEF) for the support in the acquisition and management of human specimens. AE is incumbent of the Leah Omenn Career Development Chair and is supported by research grants from the European research program (CIG618113, ERC614204), the Israel Science Foundation (1343/13; 1952/13) and from the Minerva grant award (711730). AE received additional support from the Adelis Foundation, the Henry S. and Anne S. Reich Research Fund, the Dukler Fund for Cancer Research, the Paul Sparr Foundation, the Saul and Theresa Esman Foundation, from Joseph Piko Baruch, and from the estate of Fannie Sherr. E.R. gratefully acknowledges support from I-CORE Center of Excellence in Gene Regulation in Complex Human Disease, Grant No. 41/11, the Israel Science Foundation (grant No 41/11) and support from IPST, NIST and an NIH R33 CA225291 grants to his lab. The work of A.C. is supported by the Basque Department of Industry, Tourism and Trade (EtorTek) and education (IKERTALDE I.T.1106-16), the BBVA foundation, the MINECO (SAF2016-79381-R,

FEDER/EU), the MSCA-ITN-ETN (Proposal number: 721532), the European Research Council Starting Grant (336343) and the European Research Council Proof-of-Concept program (754627 – MetaboMARKER). The participation of A.C. and M.U. as part of CIBERONC was co-funded with FEDER funds. SH and HK were funded in part by NSF award 1564785. DMW was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. Y.S is supported by the Israel Science Foundation grant number 696/17 and ERC-2016-PoC (754282).

Author Contribution

JSL and LA led the genomic analysis and molecular experiments respectively; HK and SH analyzed the proteomic and mutational data; NCa, RE performed the immunohistochemistry staining and quantifications; SR, RK, ASi, HW, SP, performed molecular experiments; LAg, NS, ASc were responsible for the animal experiments; NAu performed metabolic data analysis; QS, SCSN, AB, SM, performed the metabolic analysis; DS, DD, RE, AC, MU, provided the human cells and samples; SK, RL, EB, AA, YS performed the peptidomic analysis; NSG performed the viral experiments; DMW contributed to the thought process, ER and AE initiated and led the study and wrote the manuscript. All authors were involved in discussions about study design and reviewed the manuscript.

Declaration of Interests

The authors declare initiating a patent for Methods of diagnosing and prognosing cancer; US Provisional Patent Application No. 62/470,276.

References

- Ah Mew, N., Simpson, K.L., Gropman, A.L., Lanpher, B.C., Chapman, K.A., and Summar, M.L. (1993). Urea Cycle Disorders Overview. In GeneReviews((R)), M.P. Adam, H.H. Ardinger, R.A. Pagon, S.E. Wallace, L.J.H. Bean, H.C. Mefford, K. Stephens, A. Amemiya, and N. Ledbetter, eds. (Seattle (WA)).
- Alexandrov, L.B., Ju, Y.S., Haase, K., Van Loo, P., Martincorena, I., Nik-Zainal, S., Totoki, Y., Fujimoto, A., Nakagawa, H., Shibata, T., *et al.* (2016). Mutational signatures associated with tobacco smoking in human cancer. *Science* 354, 618-622.
- Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Borresen-Dale, A.L., *et al.* (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415-421.
- Allegra, C.J., Hoang, K., Yeh, G.C., Drake, J.C., and Baram, J. (1987). Evidence for direct inhibition of de novo purine synthesis in human MCF-7 breast cells as a principal mode of metabolic inhibition by methotrexate. *J Biol Chem* 262, 13520-13526.

Alves, G., Ogurtsov, A.Y., and Yu, Y.K. (2007). RAId_DbS: peptide identification using database searches with realistic statistics. *Biol Direct* 2, 25.

Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. *Bioinformatics* 32, 511-517.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., *et al.* (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603-607.

Beerenwinkel, N., Schwarz, R.F., Gerstung, M., and Markowetz, F. (2015). Cancer evolution: mathematical models and computational inference. *Syst Biol* 64, e1-25.

Ben-Sahra, I., Howell, J.J., Asara, J.M., and Manning, B.D. (2013). Stimulation of de novo pyrimidine synthesis by growth signaling through mTOR and S6K1. *Science* 339, 1323-1328.

Berquist, B.R., and Wilson, D.M., 3rd (2012). Pathways for repairing and tolerating the spectrum of oxidative DNA lesions. *Cancer Lett* 327, 61-72.

Brosnan, M.E., and Brosnan, J.T. (2007). Orotic acid excretion and arginine metabolism. *J Nutr* 137, 1656S-1661S.

Budhu, A., Roessler, S., Zhao, X., Yu, Z., Forgues, M., Ji, J., Karoly, E., Qin, L.X., Ye, Q.H., Jia, H.L., *et al.* (2013). Integrated metabolite and gene expression profiles identify lipid biomarkers associated with progression of hepatocellular carcinoma and patient outcomes. *Gastroenterology* 144, 1066-1075 e1061.

Cancer Genome Atlas, N. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61-70.

Chaerkady, R., Harsha, H.C., Nalli, A., Gucek, M., Vivekanandan, P., Akhtar, J., Cole, R.N., Simmers, J., Schulick, R.D., Singh, S., *et al.* (2008). A quantitative proteomic approach for identification of potential biomarkers in hepatocellular carcinoma. *Journal of proteome research* 7, 4289-4298.

Chambers, M.C., Maclean, B., Burke, R., Amodei, D., Ruderman, D.L., Neumann, S., Gatto, L., Fischer, B., Pratt, B., Egertson, J., *et al.* (2012). A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* 30, 918-920.

Chen, H., Lin, F., Xing, K., and He, X. (2015). The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nat Commun* 6, 6367.

Consortium, G.T. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet* 45, 580-585.

Cortes-Ciriano, I., Lee, S., Park, W.Y., Kim, T.M., and Park, P.J. (2017). A molecular portrait of microsatellite instability across multiple cancers. *Nat Commun* 8, 15180.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26, 1367-1372.

DeVito, S.R., Ortiz-Riano, E., Martinez-Sobrido, L., and Munger, J. (2014). Cytomegalovirus-mediated activation of pyrimidine biosynthesis drives UDP-sugar synthesis to support viral protein glycosylation. *Proc Natl Acad Sci U S A* 111, 18019-18024.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R., and Palsson, B.O. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A* 104, 1777-1782.

Eggermont, A.M., Chiarion-Sileni, V., Grob, J.J., Dummer, R., Wolchok, J.D., Schmidt, H., Hamid, O., Robert, C., Ascierto, P.A., Richards, J.M., *et al.* (2016). Prolonged Survival in Stage III Melanoma with Ipilimumab Adjuvant Therapy. *N Engl J Med* 375, 1845-1855.

Folger, O., Jerby, L., Frezza, C., Gottlieb, E., Ruppin, E., and Shlomi, T. (2011). Predicting selective drug targets in cancer through metabolic networks (vol 7, pg 501, 2011). *Molecular Systems Biology* 7.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., *et al.* (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal* 6, pii.

Gerrits, G.P., Gabreels, F.J., Monnens, L.A., De Abreu, R.A., van Raaij-Selten, B., Niezen-Koning, K.E., and Trijbels, J.M. (1993). Argininosuccinic aciduria: clinical and biochemical findings in three children with the late onset form, with special emphasis on cerebrospinal fluid findings of amino acids and pyrimidines. *Neuropediatrics* 24, 15-18.

Gubin, M.M., Artyomov, M.N., Mardis, E.R., and Schreiber, R.D. (2015). Tumor neoantigens: building a framework for personalized cancer immunotherapy. *J Clin Invest* 125, 3413-3421.

Haradhvala, N.J., Polak, P., Stojanov, P., Covington, K.R., Shinbrot, E., Hess, J.M., Rheinbay, E., Kim, J., Maruvka, Y.E., Braunstein, L.Z., *et al.* (2016). Mutational Strand Asymmetries in Cancer Genomes Reveal Mechanisms of DNA Damage and Repair. *Cell* 164, 538-549.

Hoof, I., Peters, B., Sidney, J., Pedersen, L.E., Sette, A., Lund, O., Buus, S., and Nielsen, M. (2009). NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics* 61, 1-13.

Hugo, W., Zaretsky, J.M., Sun, L., Song, C., Moreno, B.H., Hu-Lieskovan, S., Berent-Maoz, B., Pang, J., Chmielowski, B., Cherry, G., *et al.* (2016). Genomic and Transcriptomic Features of Response to Anti-PD-1 Therapy in Metastatic Melanoma. *Cell* 165, 35-44.

Ishihama, Y., Rappsilber, J., Andersen, J.S., and Mann, M. (2002). Microcolumns with self-assembled particle frits for proteomics. *J Chromatogr A* 979, 233-239.

Janin, J. (1979). Surface and inside volumes in globular proteins. *Nature* 277, 491-492.

Jinks-Robertson, S., and Bhagwat, A.S. (2014). Transcription-associated mutagenesis. *Annu Rev Genet* 48, 341-359.

Jurtz, V., Paul, S., Andreatta, M., Marcatili, P., Peters, B., and Nielsen, M. (2017). NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* 199, 3360-3368.

Kabbarah, O., Nogueira, C., Feng, B., Nazarian, R.M., Bosenberg, M., Wu, M., Scott, K.L., Kwong, L.N., Xiao, Y., Cordon-Cardo, C., *et al.* (2010). Integrative genome comparison of primary and metastatic melanomas. *PLoS One* 5, e10770.

Karosiene, E., Lundegaard, C., Lund, O., and Nielsen, M. (2012). NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. *Immunogenetics* 64, 177-186.

Kim, J., Hu, Z., Cai, L., Li, K., Choi, E., Faubert, B., Bezwada, D., Rodriguez-Canales, J., Villalobos, P., Lin, Y.F., *et al.* (2017). CPS1 maintains pyrimidine pools and DNA synthesis in KRAS/LKB1-mutant lung cancer cells. *Nature* 546, 168-172.

Kryazhinskiy, S., and Plotkin, J.B. (2008). The population genetics of dN/dS. *PLoS Genet* 4, e1000304.

Lee, Y.Y., Li, C.F., Lin, C.Y., Lee, S.W., Sheu, M.J., Lin, L.C., Chen, T.J., Wu, T.F., and Hsing, C.H. (2014). Overexpression of CPS1 is an independent negative prognosticator in rectal cancers receiving concurrent chemoradiotherapy. *Tumour biology : the journal of the International Society for Oncodevelopmental Biology and Medicine* 35, 11097-11105.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078-2079.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep* 50, 163-170.

McGranahan, N., and Swanton, C. (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* 168, 613-628.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., *et al.* (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., *et al.* (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature* 534, 55-62.

Miller, A., Asmann, Y., Cattaneo, L., Braggio, E., Keats, J., Auclair, D., Lonial, S., Network, M.C., Russell, S.J., and Stewart, A.K. (2017). High somatic mutation and neoantigen burden are correlated with decreased progression-free survival in multiple myeloma. *Blood Cancer J* 7, e612.

Milner, E., Gutter-Kapon, L., Bassani-Strenberg, M., Barnea, E., Beer, I., and Admon, A. (2013). The effect of proteasome inhibition on the generation of the human leukocyte antigen (HLA) peptidome. *Mol Cell Proteomics* 12, 1853-1864.

Musch, W., Verfaillie, L., and Decaux, G. (2006). Age-related increase in plasma urea level and decrease in fractional urea excretion: clinical application in the syndrome of inappropriate secretion of antidiuretic hormone. *Clin J Am Soc Nephrol* 1, 909-914.

Nagamani, S.C., and Erez, A. (2016). A metabolic link between the urea cycle and cancer cell proliferation. *Mol Cell Oncol* 3, e1127314.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* 3, 418-426.

Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. *Genome Med* 8, 33.

Orth, J.D., Thiele, I., and Palsson, B.O. (2010). What is flux balance analysis? *Nat Biotechnol* 28, 245-248.

Rabinovich, S., Adler, L., Yizhak, K., Sarver, A., Silberman, A., Agron, S., Stettner, N., Sun, Q., Brandis, A., Helbling, D., *et al.* (2015). Diversion of aspartate in ASS1-deficient tumours fosters de novo pyrimidine synthesis. *Nature*.

Rappsilber, J., Ishihama, Y., and Mann, M. (2003). Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* 75, 663-670.

Rasmussen, M., Fenoy, E., Harndahl, M., Kristensen, A.B., Nielsen, I.K., Nielsen, M., and Buus, S. (2016). Pan-Specific Prediction of Peptide-MHC Class I Complex Stability, a Correlate of T Cell Immunogenicity. *J Immunol* 197, 1517-1524.

Riaz, N., Havel, J.J., Makarov, V., Desrichard, A., Urba, W.J., Sims, J.S., Hodi, F.S., Martin-Algarra, S., Mandal, R., Sharfman, W.H., *et al.* (2017). Tumor and Microenvironment Evolution during Immunotherapy with Nivolumab. *Cell* 171, 934-949 e915.

Richmond, A.L., Kabi, A., Homer, C.R., Marina-Garcia, N., Nickerson, K.P., Nesvizhskii, A.I., Sreekumar, A., Chinnaiyan, A.M., Nunez, G., and McDonald, C. (2012). The nucleotide synthesis enzyme CAD inhibits NOD2 antibacterial function in human intestinal epithelial cells. *Gastroenterology* 142, 1483-1492 e1486.

Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., *et al.* (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124-128.

Robitaille, A.M., Christen, S., Shimobayashi, M., Cornu, M., Fava, L.L., Moes, S., Prescianotto-Baschong, C., Sauer, U., Jenoe, P., and Hall, M.N. (2013). Quantitative phosphoproteomics reveal mTORC1 activates de novo pyrimidine synthesis. *Science* 339, 1320-1323.

Roessler, S., Long, E.L., Budhu, A., Chen, Y., Zhao, X., Ji, J., Walker, R., Jia, H.L., Ye, Q.H., Qin, L.X., *et al.* (2012). Integrative genomic identification of genes on 8p associated with hepatocellular carcinoma progression and patient survival. *Gastroenterology* 142, 957-966 e912.

Roh, W., Chen, P.L., Reuben, A., Spencer, C.N., Prieto, P.A., Miller, J.P., Gopalakrishnan, V., Wang, F., Cooper, Z.A., Reddy, S.M., *et al.* (2017). Integrated molecular analysis of tumor biopsies on sequential CTLA-4 and PD-1 blockade reveals markers of response and resistance. *Sci Transl Med* 9.

Rosenberg, J.E., Hoffman-Censits, J., Powles, T., van der Heijden, M.S., Balar, A.V., Necchi, A., Dawson, N., O'Donnell, P.H., Balmanoukian, A., Loriot, Y., *et al.* (2016). Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *Lancet* 387, 1909-1920.

Royo, F., Zuniga-Garcia, P., Torrano, V., Loizaga, A., Sanchez-Mosquera, P., Ugalde-Olano, A., Gonzalez, E., Cortazar, A.R., Palomo, L., Fernandez-Ruiz, S., *et al.* (2016). Transcriptomic profiling of urine extracellular vesicles reveals alterations of CDH3 in prostate cancer. *Oncotarget* 7, 6835-6846.

Salerno, C., D'Eufemia, P., Celli, M., Finocchiaro, R., Crifo, C., and Giardini, O. (1999). Determination of urinary orotic acid and uracil by capillary zone electrophoresis. *J Chromatogr B Biomed Sci Appl* 734, 175-178.

Schmidt, T.T., Reyes, G., Gries, K., Ceylan, C.U., Sharma, S., Meurer, M., Knop, M., Chabes, A., and Hombauer, H. (2017). Alterations in cellular metabolism triggered by URA7 or GLN3 inactivation cause imbalanced dNTP pools and increased mutagenesis. *Proc Natl Acad Sci U S A* 114, E4442-E4451.

Shukla, S.A., Rooney, M.S., Rajasagi, M., Tiao, G., Dixon, P.M., Lawrence, M.S., Stevens, J., Lane, W.J., Dellagatta, J.L., Steelman, S., *et al.* (2015). Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat Biotechnol* 33, 1152-1158.

Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S., *et al.* (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N Engl J Med* 371, 2189-2199.

Spinelli, J.B., Yoon, H., Ringel, A.E., Jeanfavre, S., Clish, C.B., and Haigis, M.C. (2017). Metabolic recycling of ammonia via glutamate dehydrogenase supports breast cancer biomass. *Science* 358, 941-946.

Syed, N., Langer, J., Janczar, K., Singh, P., Lo Nigro, C., Lattanzio, L., Coley, H.M., Hatzimichael, E., Bomalaski, J., Szlosarek, P., *et al.* (2013). Epigenetic status of argininosuccinate synthetase and argininosuccinate lyase modulates autophagy and cell death in glioblastoma. *Cell Death Dis* 4, e458.

Terunuma, A., Putluri, N., Mishra, P., Mathe, E.A., Dorsey, T.H., Yi, M., Wallace, T.A., Issaq, H.J., Zhou, M., Killian, J.K., *et al.* (2014). MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis. *J Clin Invest* 124, 398-412.

Therneau, T.M., and Grambsch, P.M. (2000). Modeling survival data : extending the Cox model (New York: Springer).

Thompson, P.G., Smouse, P.E., Scofield, D.G., and Sork, V.L. (2014). What seeds tell us about birds: a multi-year analysis of acorn woodpecker foraging movements. *Mov Ecol* 2.

Van Allen, E.M., Miao, D., Schilling, B., Shukla, S.A., Blank, C., Zimmer, L., Sucker, A., Hillen, U., Geukes Foppen, M.H., Goldinger, S.M., *et al.* (2015). Genomic correlates of response to CTLA-4 blockade in metastatic melanoma. *Science* 350, 207-211.

van de Logt, A.E., Kluijtmans, L.A., Huigen, M.C., and Janssen, M.C. (2017). Hyperammonemia due to Adult-Onset N-Acetylglutamate Synthase Deficiency. *JIMD Rep* 31, 95-99.

van Kuilenburg, A.B., van Maldegem, B.T., Abeling, N.G., Wijburg, F.A., and Duran, M. (2006). Analysis of pyrimidine synthesis de novo intermediates in urine during crisis of a patient with ornithine transcarbamylase deficiency. *Nucleosides Nucleotides Nucleic Acids* 25, 1251-1255.

Vita, R., Overton, J.A., Greenbaum, J.A., Ponomarenko, J., Clark, J.D., Cantrell, J.R., Wheeler, D.K., Gabbard, J.L., Hix, D., Sette, A., *et al.* (2015). The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 43, D405-412.

Wei, S.C., Levine, J.H., Cogdill, A.P., Zhao, Y., Anang, N.A.S., Andrews, M.C., Sharma, P., Wang, J., Wargo, J.A., Pe'er, D., *et al.* (2017). Distinct Cellular Mechanisms Underlie Anti-CTLA-4 and Anti-PD-1 Checkpoint Blockade. *Cell* 170, 1120-1133 e1117.

Wise, D.R., and Thompson, C.B. (2010). Glutamine addiction: a new therapeutic target in cancer. *Trends Biochem Sci* 35, 427-433.

Yoshihara, K., Shahmoradgoli, M., Martinez, E., Vegesna, R., Kim, H., Torres-Garcia, W., Trevino, V., Shen, H., Laird, P.W., Levine, D.A., *et al.* (2013). Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun* 4, 2612.

Zhang, H., Lund, O., and Nielsen, M. (2009). The PickPocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to MHC-peptide binding. *Bioinformatics* 25, 1293-1299.

Main Figures

Figure 1. The expression of UC enzymes and transporters is commonly dysregulated in cancer. **A.** An illustration of the substrates channelling between the urea cycle enzymes and transporters and the pyrimidine synthesis pathway. Abbreviations: ASS1- Argininosuccinate Synthase, ASL- Argininosuccinate Lyase, OTC- Ornithine Carbamoyltransferase; CAD: Carbamoyl-Phosphate Synthetase 2 (CPS2), Aspartate Transcarbamylase (ATC), and Dihydroorotase (DHO), DHODH- Dihydroorotate Dehydrogenase, UMP Synthase- Uridine Monophosphate Synthase. **B.** Most tumor types (92%, 23 out of 25) in the TCGA database have aberrant expression of at least two components of the UC which facilitates the supply of CAD substrates (**left panel**), as compared to their expression in the corresponding normal tissues in GTEx (**right panel**). The differences remain significant vs. random choice of sets of six metabolic genes (empirical $P < 0.001$). Tumor type abbreviations: UCEC - uterine corpus endometrial carcinoma, THCA - thyroid carcinoma, TGCT - testicular germ cell tumors, STAD- stomach adenocarcinoma, SKCM - skin cutaneous melanoma, SARC - sarcoma, PRAD - prostate adenocarcinoma, PAAD - pancreatic adenocarcinoma, OV - ovarian serous cystadenocarcinoma, LUSC- lung squamous cell carcinoma; LUAD - lung adenocarcinoma, LIHC - liver hepatocellular carcinoma; LGG - brain lower grade glioma, LAML - acute myeloid leukemia, KIRP - kidney renal papillary cell carcinoma, KIRC - kidney renal clear cell carcinoma, KICH - kidney chromophobe, HNSC- head-neck squamous cell carcinoma, GBM - glioblastoma multiforme, ESCA - esophageal carcinoma, DLBC - lymphoid neoplasm diffuse large B-cell lymphoma, COAD - colon adenocarcinoma, CESC - cervical squamous cell carcinoma and endocervical adenocarcinoma, BRCA- breast invasive carcinoma, BLCA- bladder carcinoma.

Figure 2. Cancer proliferation correlates with changes in expression of UC enzymes and transporters. **A.** Tissue arrays composed of at least 10 tumors for each cancer type and 4 or more of the corresponding normal tissues were immune-stained for several UC components (scale bar

of 200 μ m). Each staining was calibrated and repeated in three technical repetitions per patient sample (OD levels were compared via a Student t-test). The quantification analyses of these plots are presented in **Figure S3A. B.** UCD scores (x-axis, equally divided into 5 bins, N~1,565 each) are positively associated with CAD expression (rank-normalized in each cancer type) in the TCGA dataset (each pair of consecutive bins was compared using a Wilcoxon rank sum test). **C.** The expression of the proliferation marker Ki-67 in the TCGA data, is significantly higher in UCD samples (UCD-score >top 45%, N=1,990) compared to UC-WT (bottom <45%, N=1,990) (Wilcoxon rank sum $P < 2.2E-16$). **D.** UCD-scores in patients' derived cells increase at more progressed cancerous states, from healthy skin (N=8) to nevi (N=9), primary melanoma (N=31), and metastatic melanoma (N=73) cells (Wilcoxon rank sum test) (Kabbarah et al., 2010). **E.** Pan cancer Kaplan-Meier (KM) survival curves (computed across all TCGA samples), showing that UCD is associated with worse survival. To confirm the robustness of the signal, we applied varying thresholds to define UCD and UC-WT samples (solid line: top- and bottom-30%, single dotted line: 45%, and double-dotted line: 15%). The statistics of the log rank test with 30% threshold are reported in the Figure.

Figure 3. Changes in nitrogenous compounds are detectable in bio-fluids of cancer patients.

A. Increased pyrimidines (uracil and thymidine) in urine of patients with prostate cancer (N=49) compared to controls (N=10), (Wilcoxon rank sum $P < 0.05$ and 0.01 , respectively). **B.** The distribution of pyrimidine to purine metabolites ratio for samples with low vs high UCD scores (top and bottom 15%); the **upper panel** shows the results for hepatocellular carcinoma (HCC) (Wilcoxon rank sum $P < 0.05$, N=9 for each group) and the **lower panel** for breast cancer (BC) tumors (Wilcoxon rank sum $P < 0.046$, N=4 for each group). **C. Upper panel** - Increased excretion of pyrimidine pathway metabolites in urine of mice with colon tumors (N=10 in each group) $P < 0.05$. **Lower panel** - A western blot demonstrating spontaneously emerging UCD in mice colon tumors compared to normal intestines (see quantification in **Figure S4B**). **D.** Mice bearing colon cancer have decreased plasma urea levels. **E.** Children with different cancers (red, N=100), have lower than normal plasma levels of urea on the day of admittance to the hospital ($P \leq 0.05$, using one sided student t-test), as compared to plasma urea levels of age and sex matched Israeli children (grey, N= 1,363,691).

Figure 4. UCD promotes genomic transversion mutations in cancer cell lines. A. Knockdown of the UC enzymes in different cancer cells increases the pyrimidine to purine ratio of synthesis favouring pyrimidines, as measured by LCMS. The figure is a representative of the mean of more than two biological repeats. ($P \leq 0.05$, one-sided paired t-test). The values of metabolites from which the ratio was calculated are presented in **Table S2** and shown as graphs in **Figure S4E. B.** Inducing perturbations in individual UC components in the directions expected to increase UCD (including OTC, ASS1, ORNT1 and citrin) significantly increases R->Y mutations (Fisher's exact test). **C.** PTMB levels after UC-perturbation in LOX and U2OS cell lines. PTMB levels after two weeks (green) are significantly higher than those after one week (yellow; bootstrap empirical $P < 0.001$ and $P < 0.03$ for LOX and U2OS, respectively).

Figure 5. UCD associates with genomic transversion mutations in tumor samples. A. Tumors with UCD have significantly higher PTMB levels versus tumors with intact UC (termed 'wild-type' (WT), Wilcoxon rank sum $P < 0.008$, N=1,990 each), while such a significant difference is not observed for transition mutations. **B.** UCD is associated with a higher level of PTMB across

different cancer types (each circle denotes the median UCD and PTMB levels in each cancer type (Spearman $R = 0.58$, $P < 0.01$). Evidently, lung cancer (LUAD, LUSC) samples are notable outliers, probably because they are confounded with a smoking mutational signature (Alexandrov et al., 2016) enriched with transversion mutations. The exclusion of these lung cancer subtypes leads to a higher correlation, $R = 0.78$ ($P < 0.007$) (see **Methods** for details). **C.** PTMB correlates with gene expression (Spearman $R = 0.8$, $P < 0.02$) in UC-perturbed cell lines but less so in the control cell lines. **D.** Tumors with UCD have significantly higher levels of PTMB at the mRNA level, based on the analysis of 18 TCGA breast cancer samples (Wilcoxon rank sum $P < 0.05$, $N = 8$ in each group). **E.** Genome wide proteomic analysis of 40 breast cancers showing that PTMB propagates to the proteome (Wilcoxon rank sum $P < 0.05$, $N = 18$ in each group). **F.** Pan cancer Kaplan-Meier survival curves showing significantly worse prognosis for patients with tumors bearing higher PTMB levels at varying thresholds for UCD and UC-WT samples (solid line 30%, single dotted line 45%, and double-dotted line 15%). The outcome of the log rank test using a 30% threshold cut-off is denoted in the Figure.

Figure 6. Higher UCD and PTMB levels are associated with increased response to immune checkpoint therapy in mouse models and patients. **A.** UCD-scores are significantly higher in responders (orange, $N = 13$) than non-responders (grey, $N = 8$) to anti-PD1 therapy (Wilcoxon rank sum $P < 0.05$). **B.** ROC curves showing higher predictive power of PTMB ($AUC = 0.77$, blue) compared to that obtained via mutational load ($AUC = 0.34$, red) in predicting the response to anti-PD1 therapy (Roh et al., 2017). In this dataset, UCD-scores could not be calculated because the gene expression of UC enzymes was not measured (due to limited coverage of nanostring measurements). **C.** Hydrophobicity (Janin, 1979) of 15 R->Y candidate neopeptides. The triangles denote the changes in the hydrophobicity of the candidate neopeptides in the UC dysregulated vs control cell lines. Among the predicted neopeptides, those that are either predicted to bind to MHC-I (Andreatta and Nielsen, 2016; Jurtz et al., 2017; Karosiene et al., 2012; Zhang et al., 2009) (Rasmussen et al., 2016) (biorxiv: <http://www.biorxiv.org/content/early/2017/08/09/174243>), or whose AA sequence is found in IEDB (Vita et al., 2015), are marked in red (see **Methods**). The four R->Y neoantigens predicted to bind to MHC-I show significantly higher hydrophobicity following R->Y mutation (paired one-sided t-test $P < 0.04$). **D.** MC-38 mouse colon cancer cells without UCD (**left**) and with UCD (**right**), were injected into C57Bl6 mice. The mice were treated intra-peritoneally with anti-PD1 immunotherapy on days 7, 12, 16, 19. Tumor volume was palpated twice a week. Cancer cells with UCD (induced *via* the knockdown of ASS1 with shASS1; (blue)) showed a significantly higher response to anti-PD1 treatment compared to the controls (yellow), as reflected by a significant decrease in tumor volume ($N = 20$ mice, 5 mice in each group, Wilcoxon rank sum $P < 0.007$). **E.** Following sacrifice on day 21, flow cytometry analysis shows significantly increased CD8 cytotoxic T cells infiltrates into the excised tumors ($N = 20$ mice, 5 mice in each group, Wilcoxon rank sum $P = 0.01$ and 0.3 , respectively for shASS1 and EV). **F.** UCD mice treated with anti-PD1 (blue) show significantly attenuated tumor growth compared to the untreated mice (yellow) ($P < 0.01$, ANOVA with Dunnett's correction).

Figure 7. Summary slide for the hypothesized “UCD effect”. While in normal tissues excess nitrogen is disposed as urea, in cancer cells most nitrogen is utilized for synthesis of macromolecules, with pyrimidine synthesis playing a major role in carcinogenesis and effecting patients' prognosis and response to ICT.

Supplemental Figures

Figure S1. Related to Figure 1-2. A specific dysregulation in expression of genes involved in UC facilitates CAD activation. **A.** A scheme showing the direction (up or down-regulation) of UC protein expression that is expected to support CAD activation (**blue arrows**) while the resulting changes in metabolites' levels following these expression alterations are represented by **red arrows**. **B. Upper panel:** Quantification of a crystal violet staining fold change in proliferation of UCD patients' fibroblasts vs normal fibroblasts controls, at 48 hours in comparison to time 0 for each cell type ($P \leq 0.05$, Student t-test). The plot is a representative of 4 biological repetitions for each cell line. **Lower panel:** Western blot demonstrating increased levels of CAD and phosphorylated CAD in specific urea cycle disorders versus normal fibroblasts. **C.** Normalized expression of *ASS1*, *SLC25A13* and *CAD* in control fibroblasts as measured by ribosome profiling following human Cytomegalovirus (CMV) infection. **D. Left panel-** Crystal violet quantification of MCF-10 non-tumorigenic epithelial cell line following UCD induction by either decreasing *ASS1* or over expressing Citrin, significantly increases proliferation. **Right panels-** RT PCR of UC genes following perturbation. **E.** Metabolic modelling predicts an association of increased nitrogen utilization with increased CAD activity, at high biomass production (that is, higher cell proliferation) conditions. **F.** Metabolic modelling predicts that following UCD, there is a significant increase in metabolic flux reactions involving pyrimidine metabolites.

Figure S2. Related to Figure 2. A specific dysregulation in expression of genes involved in UC facilitates cancer proliferation. **A.** A western blot showing the specific UC perturbations induced in different cancer cells and the resultant effect on p6K and CAD activation. **B.** Rapamycin treatment decreases CAD phosphorylation in UC perturbed cancer cells. **C. Upper left:** Quantification of crystal violet staining showing increased proliferation of different cancer cells following specific UC perturbations. **Lower left:** Rescue experiments for the UC perturbation reverses the proliferative phenotype. **Right panels:** RT PCR quantification for the changes in UC genes RNA expression levels following transfection with the specific rescue plasmid versus control plasmids. **D.** 5FU treatment decreases proliferation more significantly in UC perturbed cancer cells (**upper panels**) and restricted the enhanced pyrimidine synthesis (**lower panel**). **E. Left Panel:** *in vivo* growth of HepG2 and SKOV xenografts with UCD compared to xenografts infected with empty vector. **Right Panel:** Enhanced synthesis of labelled M+1 uracil from ^{15}N - α -glutamine in HEPG2 and SKOV perturbed cancer cells as compared to controls. **F.** Silencing of OTC in hepatocellular carcinoma cells increases the uracil/ urea ratio. For all, *P-value \leq 0.05.

Figure S3. Related to Figure 2. UCD in tumors correlates with increased tumor grade and worse prognosis. **A.** Quantification graphs for **Figure 2B** showing optical density (OD) quantifications for UC proteins compared to Proliferating cell nuclear antigen-PCNA and thyroid epithelial positive cell count. **B.** UCD-score correlates with PCNA levels (Spearman R=0.67, P<0.08). Malignant cell lines (red) show higher levels of PCNA compared to normal counterparts (blue). **C. Left panel:** Advanced thyroid tumor grade is associated with decreased expression of the UC transporter ORNT1. **Right panel-** A quantification graph for the staining performed in a blinded manner. **D.** Cox regression analysis of the UCD-score and CAD expression (in two independent univariate models and one multivariate model as competing terms) shows that both variables are independently significant.

Figure S4. Related to Figures 3. UCD increases urine pyrimidine excretion and changes nucleotide pool. **A.** Increased pyrimidine secretion relative to purines in urine of tumor bearing female (**left panel**) and male (**right panel**) mice injected with breast cancer in their mammary glands. **B.** Increased excretion of pyrimidine pathway metabolites in mice with breast and colon tumors (n=37 with tumors; n=11 controls, P<0.05, Wilcoxon ranksum test). **C.** Quantification of western blots shown in **Figure 3C (lower panel)**, demonstrating significant changes in UC

expression in colon tumors compared to normal intestine. **D.** Increased secretion of amino acids in syngeneic female and male mice bearing breast (4T1) and colon (CT26) tumors as compared to controls. **E.** Inducing UCD in MC-38 colon cancer cells decreases urine secretion of amino acids. **F.** RT-PCR of livers from breast cancer bearing mice demonstrates decreased expression of UC genes. **G.** Each graph demonstrates the concentration of each monophosphate- nucleotide in specific UCD perturbed (gray) and un-perturbed cancer cells (black). The results are demonstrated in different cancers following different UC perturbations: **upper left-** HepG2 hepatocellular carcinoma with and without OTC downregulation, **upper right-** SKOV ovarian cancer with and without SLC25A15 downregulation, **lower left-** U2OS osteosarcoma with and without ASS1 downregulation, and in **lower right** - LOX melanoma cells with and without Citrin overexpression. Following each UC perturbation in different cancers we find a consistent increase in pyrimidine synthesis compared to control, while the trend of purines levels between the UC perturbed and unperturbed controls, vary.

Figure S5. Related to Figure 4-5. Association between UCD-score with PTMB and its

functional importance. **A.** Temporal changes of UCD and PTMB levels during tumorigenesis in a mouse xenograft model. The X-axis shows the eight time points and the Y-axis shows the PTMB levels (red) and UCD-scores. **B.** UCD is not associated with mutational load in TCGA samples (Wilcoxon rank sum $P > 0.28$), where no significant difference was observed between UCD (UCD-score $>$ top 45%, $N=1,990$) and UC-WT samples (UCD-score $<$ bottom 45%, $N=1,990$) (Wilcoxon rank sum $P > 0.35$). **C.** PTMB levels are significantly greater in UCD samples (UCD-score $>$ top 45%) than UC-WT samples (UCD-score $<$ bottom 45%) when PTMB was derived from all SNPs (both synonymous and nonsynonymous; left panel, $N=1,990$ each) and non-exon SNPs (including those from intron/UTR region; right panel, $N=78$ each) (Wilcoxon rank sum $P < 0.05$). **D.** The scatter plot shows the median UCD-scores and PTMB levels for each cancer type when the smokers are removed from LUAD and LUSC cohorts (Spearman $R=0.81$, $P < 0.0001$). **E.** The scatter plot shows the median UCD-scores and PTMB levels for each cancer type when all the smokers are removed from the TCGA cohort (Spearman $R=0.57$, $P < 0.12$). After the removal, only 9 cancer types have substantial number of samples ($N > 10$). **F-G.** Gene expression correlates with PTMB level both when it is derived from (A) DNA mutation data and (B) RNA mutation data (Spearman $R=0.94$ ($P < 0.01$) and Spearman $R=0.98$ ($P < 4.9E-5$), respectively) specifically in UC (red) but not in UC-WT samples (blue). We analyzed the mutation data based on the exome-seq of 4,422 TCGA samples and the

RNAseq of 18 breast cancer samples. The genes were equally binned based on their expression levels. **H.** UCD samples (N=197) show a higher fraction of nonsynonymous R->Y transversion mutations compared to UC-intact samples (N=197) across all TCGA data (Wilcoxon rank sum $P < 4.93E-3$). Such a significant bias is not present for any of the other transversion mutation types (Y->Y, R->R, and Y->R).

Figure S6. Related to Figure 6. Impact of UCD and PTMB on ICT response and HLA-peptide presentation. **A.** UCD-scores are higher in responders (orange, N=7) compared to non-responders (grey, N=26, respectively) to anti-CTLA4 therapy. Wilcoxon rank sum $P < 0.05$ when UCD-score was calculated with RPKM-normalized RNAseq data (**left panel**) and Wilcoxon rank sum $P < 0.22$ when UCD-score was calculated with TPM-normalized data (**right panel**). **B.** Expression of CAD is less associated with ICT response than UCD both in anti-PD1 (Hugo et al., 2016) (**left panel**) and anti-CTLA4 (Van Allen et al., 2015) (**right panel**) cohort (Wilcoxon ranksum $P = 0.71$ and 0.45 , respectively). The number of responders and non-responders in each dataset is provided in **Table S5**. **C.** Peptidomics analysis shows UCD cell lines (blue, N=10,830) have higher LC-MS intensity than control cell lines (red, N=10,794) (Wilcoxon ranksum $P < 0.001$). **D.** Peptides in UC-perturbed cell lines (blue, N=10,830) show significantly higher hydrophobicity than unperturbed cell lines (red, N=10,794) (Wilcoxon $P < 7.00E-4$). **E. Left panel-** A western blot of MC-38 mouse colon cancer cells infected with different *shASS1* clones, demonstrate downregulation of ASS1 at the protein level as compared to control cells infected with an empty vector (EV). Quantification of the western blots is shown in **right panel**. **F.** RT PCR quantification shows decreased ASS1 RNA levels in MC38 infected with different *shASS1* clones as compared to MC38 infected with EV. **G.** *In vivo* tumor growth was enhanced in MC38 with *shASS1* as compared to the growth of MC38- EV tumors. **H.** Following sacrifice on day 22, flow cytometry analysis shows no significant CD4 helper T cells infiltrates between the excised UC perturbed and un perturbed tumors (N=20 mice, 5 mice in each group, Wilcoxon ranksum $P > 0.4$ both for *shASS1* and EV). **I.** *In vivo* tumor growth in mice with (red) and without (blue) anti-PD1 treatment in the control group (empty vector) (ANOVA $P > 0.12$).

STAR Methods

Main Figures	16
Supplemental Figures	19
CONTACT FOR REAGENTS AND RESOURCE SHARING	24
EXPERIMENTAL MODEL AND SUBJECT DETAILS	24
METHOD DETAILS	25
Deregulation of urea cycle components in TCGA samples	25
Joint transcriptomic and metabolomics analysis of tumor samples	25
Urea cycle dysregulation in the progression of melanoma	26
TCGA whole exome-seq analysis	26
TCGA patient survival analysis	28
Selective advantage of purine-to-pyrimidine mutation in UC dysregulated tumor	29
Detecting somatic mutations from exome-seq and RNAseq data	30
Detecting somatic mutations from proteomics data	31
Dynamics progression of PTMB after inducing UC-dysregulation	32
Genome-scale metabolic network modeling	34
Association of UCD/PTMB and the success of immune checkpoint therapy	35
Purine to pyrimidine mutation rate expected from the codon table	36
Production and purification of membrane HLA molecules	36
Identification of eluted HLA peptides	36
Identifying candidate neoantigens from HLA-peptidomics data	37
Cell cultures	38
Crystal violet staining	39
5FU and Rapamycin treatment	39
Western blotting	39
Immunohistochemistry	40
Virus infection	41
Patient samples	41
RNA processing and quantitative PCR	41
Transient transfection	41
Overexpression	42
Metabolomics analysis	42
Nucleotides analysis	43
Extraction of polar metabolites from Urine and Plasma	43
In vivo experiments	45
QUANTIFICATION AND STATISTICAL ANALYSIS	46
Statistical an analysis	46
DATA AND SOFTWARE AVAILABILITY	47
ADDITIONAL RESOURCES	47

CONTACT FOR REAGENTS AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Ayelet Erez (ayelet.erez@weizmann.ac.il). All information regarding computational analysis should be directed to the co-lead contact, Eytan Ruppin (eytan.ruppin@nih.gov).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

***In vivo* animal studies-** Animal experiments were approved by the Weizmann Institute Animal Care and Use Committee Following US National Institute of Health, European Commission and the Israeli guidelines (IACUC 21131015-4). To generate syngeneic mouse cancer models, 8 weeks old C57BL/6, SCID or BALB/c male and female mice were purchased from Envigo and randomly assigned to experimental groups.

Human studies- All patients' urine samples were obtained upon informed consent and with evaluation and approval from the corresponding ethics committee (CEIC code OHEUN11-12 and OHEUN14-14 (Royo et al., 2016)). Patients included in the study were men diagnosed with prostate adenocarcinoma and the criteria for inclusion was to be scheduled for surgery as anticancer treatment. Samples were collected between 2012 and 2016.

Urea analysis: To estimate age-stratified Urea background levels we pooled data from X=1,363,691 patients in the de-identified Clalit Health Services electronic health record (code 0194-17-COM2), and y=100 pediatric de-identified cancer patients on their day of admission to the Pediatric Hemato-Oncology Department at Souraski Medical Center (code 0016-17).

Cell lines- SKOV, U2OS and HepG2 cell lines was purchased from the ATTC. LOX IMVI was purchased from NCI60 (RRID:CVCL1381). MC38 cell line, derived from mouse colon adenocarcinoma (kindly provided by Dr. Eran Elinav Department of Immunology, The Weizmann Institute of Science). OTC and CPS1 deficient fibroblasts were purchased from Coriell Institute for Medical Research. Cells were cultured using standard procedures in a 37 °C humidified incubator with 5% CO₂ in Dulbecco's Modified Eagle's Medium (DMEM) (gibco) or RPMI 1640 medium (gibco) supplemented with 10-20% heat-inactivated fetal bovine serum, 10% pen-strep and 2 mM glutamine. MCF10A and its medium was kindly provided by Professor Yossi Yarden, Department of Biological Regulation, Weizmann Institute of Science. All cells

were tested routinely for Mycoplasma using Mycoplasma EZ-PCR test kit (#20-700-20, Biological Industries, Kibbutz Beit Ha'emek).

METHOD DETAILS

Deregulation of urea cycle components in TCGA samples

We downloaded TCGA gene expression profiles of 7,823 patients (4,723 GTEx healthy tissue samples) encompassing 25 cancer types for which there is corresponding healthy control samples via UCSC Xena browser (<http://xena.ucsc.edu>) (**Supplemental Data**). We compared the expression of the 6 genes involved in urea cycle (*ASL*, *ASS1*, *CPS1*, *OTC*, *SLC25A13*, and *SLC25A15*) in these cancer *vs* healthy tissue samples using Student's t test. Arginase was excluded from the UCD score as its addition did not show any effect on all evaluated parameters (**Data not shown**). Components with significant fold changes in specific tumor types are presented in **Figure 1B**. Urea cycle deregulation is a result of coordinated alterations in urea cycle enzyme activities, where *CPS1* and *SLC25A13* tend to be up-regulated, while *ASL*, *ASS1*, *OTC* and *SLC25A15* tend to be down-regulated to increase substrate supply to CAD and enhance pyrimidine synthesis. For each sample, we then determine its *UC genes dysregulation score (UCD-score)*, which is a weighted sum of the expression of 6 UC genes, where 1 or -1 was assigned as weights depending on the direction implied by the UCD signature defined above (thus it can range from a positive to a negative value), i.e.

$$\text{UCD - score} = -ASL - ASS1 + CPS1 - OTC + SLC25A13 - SLC25A15,$$

where the names of genes denote their gene expression levels. We divided the tumor samples equally into 5 bins based on the UCD-score and compared the CAD expression across these bins (where the CAD expression is rank-normalized in each cancer type to control for cancer types) using a Wilcoxon rank sum test.

Joint transcriptomic and metabolomics analysis of tumor samples

We analyzed recently published data of joint transcriptomic and metabolomic measurements across 58 breast cancer (BC) tumors *vs* controls (Terunuma et al., 2014) and 29 such samples in hepatocellular carcinoma (HCC) (Roessler et al., 2012) to further study the association between UCD and metabolites levels in clinical samples. For each sample, we computed a score denoting

the ratio of pyrimidine to purine metabolite levels in the given sample, we then divided the samples into two groups based on their UCD scores and performed Wilcoxon rank sum test comparing the two groups, in each of these two cancer types.

Urea cycle dysregulation in the progression of melanoma

We analyzed gene expression data of four types of skin samples from human subjects, namely normal skin (N=8), nevi (N=9), primary (N=31) and metastatic tumor samples (N=73) (Kabbarah et al., 2010). UCD-score was calculated in each sample and compared between the four distinct types of samples in the order of progression using Wilcoxon rank sum test. In comparing the primary and metastatic tumor samples, we controlled for patient age through multi-class rank sum test using R library ‘nestedRanksTest’ (Thompson et al., 2014).

TCGA whole exome-seq analysis

We downloaded TCGA mutation profiles for 4,858 tumors encompassing 16 cancer types from cbiportal (Gao et al., 2013) on Feb 1, 2017. For mutation analysis, we used the data from cbiportal as it integrates the mutation analysis from different TCGA centers to avoid center specific bias in mutation calls. (For the analyses described in the Methods Section “*Deregulation of urea cycle components in TCGA patient samples,*” we used the TCGA data from UCSC Xena browser (<http://xena.ucsc.edu>) because it includes TCGA RNAseq data normalized together with GTEx healthy tissue samples, which are absent in cbiportal.) We focused on 4,422 samples that have more than 20 point mutation events (where sufficient R->Y and Y->R nonsynonymous mutation event is expected) since our analysis (as described below) requires sufficient number of mutation events in each sample. This results in 758,282 single point mutation events (including 535,296 non-synonymous mutations) in 16 cancer types (**Supplemental Data**).

To study the *Pyrimidine-rich Transversion Mutational Bias (PTMB)* we consider the fraction of transversions from puRines (R) to pYrimidines (Y), $f(R \rightarrow Y)$ denotes the ratio of R->Y point mutations to all point mutations on the DNA sense strand in a given sample. The fraction of transversions from pyrimidines to purines (Y->R), is defined in an analogous manner. PTMB is defined as the difference between the two fractions, i.e.

$$PTMB = \frac{N(R \rightarrow Y) - N(Y \rightarrow R)}{\text{mutational load}} = f(R \rightarrow Y) - f(Y \rightarrow R),$$

where $N(R \rightarrow Y)$ and $N(Y \rightarrow R)$ denote the number of R->Y and Y->R single nucleotide polymorphisms (SNPs) on DNA sense strand, respectively, and ‘mutational load’ is the total

number of SNPs in a given sample. PTMB was calculated using nonsynonymous SNPs unless explicitly denoted otherwise. A sample was marked to be PTMB-biased if its PTMB level is greater than zero, while a sample with significant UC dysregulation were determined if its UCD-score is greater than the median rank-normalized UCD-score of the corresponding healthy tissue (**Table S1**).

We analyzed the association between UCD and R->Y transversion mutations using three different approaches: (1) we compared PTMB in UC dysregulated samples (UC-Dys; top 45% UCD scores) vs UC intact samples (UC-WT; bottom 45% UCD scores) at the pancancer level using a Wilcoxon rank sum test. We considered PTMB derived from (i) nonsynonymous SNPs, (ii) all SNPs (including both synonymous and nonsynonymous), and (iii) non-exon SNPs (including introns, UTR, intergenic region, splice sites and transcription start sites). For non-exon SNPs, we considered 175 TCGA samples having more than 20 non-exon mutations in each sample. We confirmed that the association between UCD and PTMB is not a mere consequence of UCD samples carrying more purines on the original sense strand (or UC-intact samples carrying more pyrimidines on the original sense strand) (Wilcoxon rank sum $P > 0.3$).

We have examined the UCD/PTMB association while controlling for each of the 30 mutational signatures (<http://cancer.sanger.ac.uk/cosmic/signatures>), one at a time, using an extension of Wilcoxon rank sum test for multi-class data using R library 'nestedRanksTest' (Thompson et al., 2014). For each signature, we divided the samples into two classes depending on the exposure of the mutational signature in the given sample (using median contribution level as threshold). We then performed a multi-class Wilcoxon rank sum test to test for the association of UCD-score with PTMB, taking into account different cancer types (FDR-corrected $P < 0.05$ for all mutational signatures).

To make sure UCD itself is not associated with the known mutational signatures, we evaluated the association between UCD-score and the exposure to 30 mutational signatures in each cancer type. When only the mutational signatures enriched with transversion mutations were considered in each cancer type (Mutational Signatures 3, 4, 5, 6, 8, 9, 10, 13, 14, 17, 18, 20, 28, 29; <http://cancer.sanger.ac.uk/cosmic/signatures>), we found that UCD significantly correlates with mutational signature 5 and 6 in pancreatic cancer (Spearman $R = 0.88$ for both, $FDR < 0.007$ and 0.01 , respectively), constituting a very small fraction of the space of possible associations. (2) We analyzed the correlation across cancer types between median UCD-scores and median PTMB

levels of each cancer type. To evaluate the potential confounding effect of smoking (whose mutational signature is enriched with transversion mutation), we obtained the smoking annotation of TCGA samples from (Alexandrov et al., 2016). We repeated the same analysis while removing (i) two lung cancer types (LUAD and LUSC), (ii) smokers in the two lung cancer types and (iii) all smoker samples. We checked whether there is a significant *increase* in the correlation between the UCD and PTMB scores after removal (e.g. **Figure S5C**) compared to that before removal (**Figure 5B**). When we remove all smokers in the analysis, we are left with sufficient number ($N > 10$) of samples for 9 out of 16 cancer types (**Figure S5E**) (nevertheless, we still observe a strong correlation (though with marginal significance) between UCD-scores and PTMB across cancer types). (3) Among genes with more than 10 mutation events in TCGA data, we identified cancer drivers whose PTMB levels are specifically associated with UCD using Wilcoxon rank sum test.

TCGA patient survival analysis

We performed two different analyses to identify the association of UCD-score with patient survival: Kaplan Meier analysis and Cox proportional hazard model. We compared the survival of patients with high-UCD score (top 15, 30, 45-percentile) and low-UCD score (bottom 15, 30, 45-percentile) using logrank test (Mantel, 1966), and the effect size were quantified by the difference in the area under the curves (ΔAUC). To control for potential confounders, we performed a Cox regression analysis, while controlling for patients' age, sex, race, and cancer types.

$$h_s(t, \text{patient}) \sim h_{0s}(t) \exp(\beta_{UCD} UCD + \beta_{age} \text{age} + \beta_p \text{purity}), \quad (2)$$

where s is an indicator variable over all possible combinations of patients' stratifications based on race, sex and cancer types. h_s is the hazard function (defined as the risk of death of patients per time unit), and $h_{0s}(t)$ is the baseline-hazard function at time t of the s^{th} stratification. The model contains three covariates: (i) *UCD*: UCD-score calculated from the urea cycle deregulation signatures (where the gene expression data was Z-score transformed in reference to normal or all other tumor samples (Gao et al., 2013) and quantile normalized), (ii) *age*: age of the patient at the time of diagnosis, and (iii) *purity*: the purity of tumor biopsy sample (Yoshihara et al., 2013). The β s are the regression coefficients of the covariates, which quantify the effect of covariates survival,

determined by standard likelihood maximization of the model (Therneau and Grambsch, 2000). For this analysis, we used TCGA cBioportal data described in Methods Section “TCGA whole exome-seq analysis.”

We performed an analogous analysis to identify the association between patient survival and PTMB (derived from all SNPs) while considering additional confounders such as microsatellite instability (MSI) (Cortes-Ciriano et al., 2017), known COSMIC mutational signatures (those associated with transversion mutations - mutational signatures 3, 4, 5, 6, 8, 9, 10, 13, 14, 17, 18, 20, 28, 29) (Alexandrov et al., 2013), and smoking annotations (Alexandrov et al., 2016).

MSI profile (total MSI events) is obtained from (Cortes-Ciriano et al., 2017), the exposure of 30 COSMIC mutational signatures from (<http://cancer.sanger.ac.uk/cosmic/signatures>) and the smoking information (based on the packs smoked per year) from (Alexandrov et al., 2013). We considered 1,033 out of the original 4,422 samples where we have all necessary information available. The log-likelihood ratios of the Cox models with and without PTMB while controlling for all the covariates listed above shows significant improvement of goodness of fit (log-likelihood ratio test $P < 0.05$). The data used for this analysis is deposited in **Supplemental Data**.

Selective advantage of purine-to-pyrimidine mutation in UC dysregulated tumor

We assessed the strength of selection (dN/dS), the ratio between the rate of nonsynonymous substitutions (dN) and the rate of synonymous substitutions (dS) for different types of mutations. Generally, dN and dS are computed in two steps: (i) assessing the number of nonsynonymous substitutions (N) per nonsynonymous site ($pN = N/n$; n the number of N sites in the compared region) and the number of synonymous substitutions (S) per synonymous site ($pS = S/s$; s the number of S sites in the compared region), and (ii) applying methods that transform pN to dN and pS to dS , accounting for the possibility that a given site is mutated more than once (e.g., (Nei and Gojobori, 1986)). However, because during tumor evolution the probability that a single site is mutated more than once is low (Beerenwinkel et al., 2015; Kryazhimskiy and Plotkin, 2008), we approximate the rates dN , dS by pN and pS , respectively.

Specifically, to assess the selection advantage (SA) of R->Y mutation relative to all other types of mutations (i.e., R->R, Y->R and Y->Y), we used the following formula:

$$SA = \frac{\left(\frac{dN}{dS}\right)_{R \rightarrow Y}}{\left(\frac{dN}{dS}\right)_{others}} = \frac{\left(\frac{pN}{pS}\right)_{R \rightarrow Y}}{\left(\frac{pN}{pS}\right)_{others}} = \frac{N_{R \rightarrow Y}}{S_{R \rightarrow Y}} \bigg/ \frac{N_{others}}{S_{others}}, \quad (1)$$

where $(dN/dS)_{R \rightarrow Y}$ and $(dN/dS)_{others}$ denote the selection of R->Y mutations and of all other types of mutations, respectively. We then compared the selection advantage of R->Y mutation specific to UCD by calculating the ratio between selection advantages of UC dysregulated vs UC intact samples. We considered TCGA samples that have at least 20 mutations (to focus on the samples with sufficient nonsynonymous ($N > 15$) and synonymous ($N > 5$) SNPs), leading to 1,313 samples in 16 cancer types.

Detecting somatic mutations from whole exome-seq and RNAseq data

To capture variants in the coding region, we downloaded exome-seq data of 18 individual breast cancer and matched normal samples from TCGA portal. For each read-alignment (BAM) file of normal and cancer we called variants using the GATK (V. 3.6) '*HaplotypeCaller*' (Li et al., 2009; McKenna et al., 2010) utility with same hg38 assembly that the TCGA used for exome-seq mapping and applying '*-ERC GVCF*' mode to produce a comprehensive record of genotype likelihoods for every position in the genome regardless of whether a variant was detected at that site or not. The goal of using the GVCF mode was to capture confidence score for every site represented in a paired normal and cancer cohort for calling somatic mutation in cancer. Next we combined the paired GVCFs from each paired cohorts using GATK's '*GenotypeGVCFs*' utility yielding genotype likelihood scores for every variant in cancer and the paired normal sample. Next, we used GATK's '*VariantRecalibrator*' utility using dbSNP VCF (v146:

ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b146_GRCh38p2/VCF) file by selecting annotation criteria of QD;MQ;MQRankSum;ReadPosRankSum;FS;SOR, followed by GATK's '*ApplyRecalibration*' utility with '*SNP*' mode. Next, using GATK's '*VariantFiltration*' utility we selected the variants with $VQSLOD \geq 4.0$. Finally, somatic mutations were defined as the loci whose genotype (1/1, 0/1, or 0/0 with '*PL*' (Phred-scaled likelihood of the genotype) score = 0, i.e., highest confidence) in cancer is distinct from that in paired normal. The final somatic mutations were mapped on an exonic site of a transcript by '*bcftools*' tool (V. 1.3) (Li et al., 2009) using BED file of coding region in hg38 assembly.

To call variants in RNA, we downloaded BAM files of RNA-Seq data for the same normal and cancer cohorts as above. First we used GATK's '*SplitNCigarReads*' utility to split the reads into

exon segments and hard-clipped to any sequence overhanging into the intronic regions. Next, we used GATK's '*HaplotypeCaller*' utility using the same hg38 assembly that the TCGA used for RNA-Seq mapping. To reduce false positive and false negative calls we used '*dontUseSoftClippedBases*' argument with the '*HaplotypeCaller*' with minimum phred-scaled confidence threshold for calling variants set to be 20. We then filtered the variants using '*VariantFiltration*' utility based on Fisher Strand values ($FS > 30$) and Qual By Depth values ($QD < 2.0$). Finally, we used each of the output VCF files for annotation of coding regions on the transcripts to which the variants were mapped by using '*bcftools*' with BED file of coding region in hg38 assembly (**Supplemental Data**). Based on this data, we compared the overall R->Y mutation bias, $f(R->Y)-f(Y->R)$, in UC dysregulated (UCD-score >top 45%) vs UC intact (UCD-score <bottom 45%) samples using a Wilcoxon rank sum test. The same procedure was applied to identify the frequency of transversion mutations in cell lines with UC dysregulation followed by a Fisher's exact test of R->Y versus Y->R mutation with background as their expected frequencies from CCLE (Barretina et al., 2012) mutation data (file '*CCLE_hybrid_capture1650_hg19_NoCommonSNPs_NoNeutralVariants_CDS_2012.05.07.maf*', N=905). The same pipeline was used to analyze the exome-seq and RNAseq data from cell line experiments with human reference genome hg19 human reference genome.

Detecting somatic mutations from proteomics data

To map the DNA variants to protein sequence, we downloaded peptide spectrum (PSM) data for 40 breast cancer samples, out of which only 4 samples overlapped with the samples analyzed for DNA mutations calls above. For each transcript in the somatic variant VCF file, we constructed complete coding sequence of RNA using GATK's '*FastaAlternateReferenceMaker*' utility. On this variant incorporated coding sequence, we captured codon that is affected by this variant site and *in silico* translated it into an amino acid; meanwhile, if the translated amino acid differs from reference amino acid we call it as '*non-synonymous*' change and otherwise '*synonymous*' (**Supplemental Data**). Based on this data, we compared the overall R->Y mutation-mapped amino acid changes in UC dysregulated (UCD-score >top 45%) vs UC intact (UCD-score <bottom 45%) samples using a Wilcoxon rank sum test.

Dynamics progression of PTMB after inducing UC-dysregulation

First, UCD-score and PTMB levels were calculated from the gene expression and mutation profiles from the mouse xenograft study where the evolutionary history of a mouse xenograft model of *HRAS*-mutated MCF10A was followed at eight different time points (Chen et al., 2015). Since UCD levels increase initially (time point 1 is significantly lower than the UCD values in other points) and then stays at about similar levels throughout, we checked whether UCD-score at time point 1 lies far off from the distribution of UCD-score at all other time points ($P < 0.016$) and performed a correlation analysis between time course and PTMB levels.

Second, we knocked-down *ASS1* in U2OS osteosarcoma cells using sh*ASS1* and overexpressed citrin using over expressing vector in LOX melanoma cell-lines. We sequenced the genome at two time points (1 week and 2 weeks following UCD induction), and compared their PTMB levels. Dragen was used to carry out germline variant calling for these samples. All most all the called variants (>99%) are biallelic. Dragen filters the variants based on some hard filter criteria. For SNPs, the filter criterion was $QD < 2.0 \parallel MQ < 30.0 \parallel FS > 60.0 \parallel MQRankSum < -12.5 \parallel ReadPosRankSum < -8.0$. For indel, the filter criterion was $QD < 2.0 \parallel ReadPosRankSum < -20.0 \parallel FS > 200.0$. After filtering based on the above criteria, the VQSR was performed over the variants. We compared their PTMB levels, where empirical p-values were calculated by bootstrapping of 1,000 times.

Potential mechanism of PTMB

To understand the potential mechanism by which pyrimidine-rich nucleotide imbalance induces PTMB, we investigated whether there is a correlation between PTMB and gene expression when UC enzymes are dysregulated both via causal cell line experiments and via mining TCGA patient data. To this end we have induced UCD in 4 different cancer cell lines by perturbing UC enzymes (in the same manner as described **Figure 4B**). After the induction of UCD, we checked that the perturbed cell lines show significant correlation between PTMB and gene expression (**Figure 5C**), while we confirmed that the UC-unperturbed cell lines do not. Furthermore, we have checked whether PTMB correlates with corresponding expression levels across TCGA UC-dysregulated samples both at DNA and RNA level, in which we binned the genes based on their expression levels and PTMB and median expression of each bin was considered for correlation analysis (**Figure S5F-G**).

Our analysis can be summarized that the pyrimidine-rich nucleotide imbalance (i) preferentially induces mutation on the DNA sense strand, (ii) biases mutation patterns from purine to pyrimidine resulting in the PTMB signature, and (iii) this type of mutational bias correlates with transcription levels. This cannot be explained via transcription-coupled repair, where the expression level is expected to anti-correlate with mutation rate. The closest explanation found in the literature is transcription-coupled damage, where the sense strand, which is left unprotected during transcription, accumulates mutations (Haradhvala NJ, Cell 2016). While the precise mechanism underlying PTMB requires further elucidation, a hypothetical model could be constructed based on current knowledge.

Genes undergoing high expression have been shown to be prone to mutagenic events (Jinks-Robertson and Bhagwat, 2014). This characteristic (so-called transcription-associated mutagenesis) likely stems from the fact that actively transcribed regions are more susceptible to hydrolytic decay or reactions with endogenous chemical species, such as oxygen radicals, due to the single-stranded nature of the transcription bubble. The sense strand of DNA is particularly at risk for damage accumulation, as it is unprotected by the nascent RNA transcript that hybridizes to the anti-sense (template) strand, the transcriptional apparatus or transcription-coupled repair that preferentially removes blocking lesions from the transcribed strand. While this feature is consistent with the mutagenic bias we observe (a) in regions of high gene expression and (b) on the sense strand, it does not by itself explain PTMB in the context of an imbalanced nucleotide pool.

To explain how PTMB arises on this background, we further hypothesize that the increased level of DNA damage observed in highly transcribed regions is further processed for repair in an error-prone manner affected by the nucleotide imbalance. That is, presuming that most of the DNA damage generated would be through endogenous processes, such as spontaneous hydrolysis or oxidative reactions, the resulting modifications (i.e., uracil, abasic sites, 8-oxoguanine, cyclopurines, etc.) would call into action either base excision or nucleotide excision repair mechanisms (Berquist and Wilson, 2012). Both repair systems rely on a re-synthesis step after damage removal, which due to an imbalance in the nucleotide pool, would be more prone to

nucleotide mis-incorporation. The consequent mispairs might evade additional repair responses, namely mismatch repair, due to the robust replicative nature of cancer cells, although mismatch repair also relies on a re-synthesis step that would presumably be prone to incorporation errors as well. Thus, in short, increased transcription would lead to increased DNA damage, which would be processed incorrectly due to error-prone DNA repair events stemming from nucleotide pool imbalance, eventually leading to mutagenesis within actively transcribed genes on the sense strand following chromosome duplication.

Genome-scale metabolic network modeling

We used genome-scale metabolic modeling to study the stoichiometric balance of nitrogen metabolism between urea production and pyrimidine synthesis. For a metabolic network with m metabolites and n reactions, the stoichiometric constraints can be represented by a stoichiometric matrix S ,

$$\sum_j S_{ij}v_j = 0, \quad (3)$$

where the entry S_{ij} represents the stoichiometric coefficients of metabolite i in reaction j , and v_j stands for the metabolic flux vector for all reactions in the model. The model assumes steady metabolic state as represented in equation (3), constraining the production rate of each metabolite to be equal to its consumption rate. In addition to the mass balance, a constraint-based model limits the space of possible fluxes in the metabolic network's reactions through a set of (in) equalities imposed by thermodynamic constraints, substrate availability, and the maximum reaction rates supported by the catalyzing enzymes and transporting proteins,

$$\alpha_j \leq v_j \leq \beta_j, \quad (4)$$

where α_j and β_j defines the lower and upper bounds of the metabolic fluxes for different types of metabolic fluxes. (i) The exchange fluxes model the metabolite exchange of a cell with the surrounding environment via transport reactions, enabling a pre-defined set of metabolites to be either taken up or secreted from the growth media. (ii) Enzymatic directionality and flux capacity constraints define lower and upper bounds on the fluxes as represented in equation (4). We used the a human metabolic network model (Duarte et al., 2007) with biomass function introduced in (Folger et al., 2011) under the Roswell Park Memorial Institute medium.

To study the metabolic alterations occurring in UC dysregulated cancer cells (having increased growth and biomass production rates, and increase CAD activity versus healthy cells), we

performed a flux-balance-based analysis (Orth et al., 2010). We computed the nitrogen utilization by subtracting the total amount of nitrogen excreted from the amount of nitrogen up taken, while taking into account the nitrogen's stoichiometry in all nitrogen-containing metabolites. We gradually increased the demand constraints for biomass production rates and the flux via the three enzymatic reactions of CAD – Carbamoyl-phosphate synthetase 2 (CPS2), Aspartate transcarbamylase (ATC), and Dihydroorotase - up to their maximal feasible values in the model.

Association of UCD/PTMB and the success of immune checkpoint therapy

Three different melanoma ICT datasets (Van Allen et al., 2015) (Hugo et al., 2016) (Roh et al., 2017), treated with anti-CTLA4 therapy and anti-PD1 therapy were analyzed. The third dataset (Roh et al., 2017) includes both anti-CTLA4 and anti-PD1, and we focused on the anti-PD1 arm because it includes a sufficient number of samples. We followed the definition of responders determined by RECIST criteria (treating complete response (CR) and partial response (PR) as responders and the progressive disease (PD) as non-responders).

UCD-score between responders and non-responders were compared in two datasets (Van Allen et al., 2015),(Hugo et al., 2016) where the expression levels of UC genes are available using Wilcoxon rank sum test. For the first dataset (Van Allen et al., 2015), we performed the analysis using both RPKM- and TPM-normalized RNAseq data; the third dataset (Roh et al., 2017) has nanostring data, where not all of the expression of 6 UC genes are available. The gene expression was rank-normalized per individual sample to calculate UCD-score. The association of CAD expression and ICT response was evaluated in an analogous manner with Wilcoxon rank sum test. While the association between mutational burden and ICT response was observed in some datasets (Rizvi et al., 2015; Rosenberg et al., 2016; Snyder et al., 2014), there are quite a few datasets (Hugo et al., 2016) (Roh et al., 2017) (Riaz et al., 2017) in which the authors report that no such association was observed. To characterize the predictive power of mutational load and compared it with that of UCD and PTMB, we evaluated their prediction accuracy in terms of area-under-the-curve (AUC) of the receiver operating characteristics (ROC) curves in the three datasets (**Table S5**).

Purine to pyrimidine mutation rate expected from the codon table

Using the codon table, we considered the 4 mutation types with respect to R/Y (i.e., R->Y, R->R, Y->Y, Y->R) and the 4 types of amino acid (AA) changes with respect to hydrophobicity (N->H, H->H, N->N, H->N), where H denotes hydrophobic AA and N denotes non-hydrophobic AA. We counted each of the cases (4 by 4 contingency table) at the first, second, and third loci of each codon. The resulting (4 by 12) table is presented as **Table S8**. We calculated the enrichment of R->Y mutation in N->H AA changes using a Fisher's exact test.

Production and purification of membrane HLA molecules

Cell line pellets were collected from 2×10^8 cells. The solution was mixed by gentle rotation in the cold for one hour with lysis buffer containing 0.25% sodium deoxycholate, 0.2mM iodoacetamide, 1mM EDTA, 1:300 Protease Inhibitors Cocktail (Sigma-Aldrich, P8340), 1mM PMSF and 1% octyl- β -D glucopyranoside in PBS. Samples were then incubated at 4°C for 1 hour. The lysates were cleared by centrifugation at 48,000g for 60 minutes at 4°C, and then were passed through a pre-clearing column containing Protein-A Sepharose beads.

HLA-I molecules were immunoaffinity purified from cleared lysate with the pan-HLA-I antibody (W6/32 antibody purified from HB95 hybridoma cells) covalently bound to Protein-A Sepharose beads. Affinity column was washed first with 10 column volumes of 400mM NaCl, 20mM Tris-HCl and then with 10 volumes of 20mM Tris-HCl, pH 8.0. The HLA peptides and HLA molecules were then eluted with 1% trifluoroacetic acid followed by separation of the peptides from the proteins by binding the eluted fraction to disposable reversed-phase C18 columns (Harvard Apparatus). Elution of the peptides was done with 30% acetonitrile in 0.1% trifluoroacetic acid (Milner et al., 2013). The eluted peptides were then cleaned using C18 stage tips as described previously (Rappsilber et al., 2003).

Identification of eluted HLA peptides

The HLA peptides were dried by vacuum centrifugation, solubilized with 0.1% formic acid, and resolved on capillary reversed phase chromatography on 0.075x300 mm laser-pulled capillaries, self-packed with C18 reversed-phase 3.5 μ m beads (Reprosil-C18-Aqua, Dr. Maisch GmbH, Ammerbuch-Entringen, Germany) (Ishihama et al., 2002). Chromatography was performed with the UltiMate 3000 RSLCnano-capillary UHPLC system (Thermo Fisher Scientific), which was coupled by electrospray to tandem mass spectrometry on Q-Exactive-Plus (Thermo Fisher Scientific). The HLA peptides were eluted with a linear gradient over 2 hours from 5 to 28%

acetonitrile with 0.1% formic acid at a flow rate of 0.15 μ l/minute. Data was acquired using a data-dependent “top 10” method, fragmenting the peptides by higher-energy collisional dissociation. Full scan MS spectra were acquired at a resolution of 70,000 at 200 m/z with a target value of 3x10⁶ ions. Ions accumulated to an AGC target value of 10⁵ with a maximum injection time of generally 100 milliseconds. The peptide match option was set to Preferred. Normalized collision energy was set to 25% and MS/MS resolution was 17,500 at 200 m/z. Fragmented m/z values were dynamically excluded from further selection for 20 seconds. The MS data were analyzed using MaxQuant (Cox and Mann, 2008) version 1.5.3.8, with 5% false discovery rate (FDR). Peptides were searched against the UniProt human database (July 2015) and customized reference databases that contained the mutated sequences identified in the sample by WES. N-terminal acetylation (42.010565 Da) and methionine oxidation (15.994915 Da) were set as variable modifications. Enzyme specificity was set as unspecific and peptides FDR was set to 0.05. The match between runs option was enabled to allow matching of identifications across the samples belonging the same patient.

HLA typing was determined from the WES data by POLYSOLVER version 1.0 (Shukla et al., 2015), and the HLA allele to which the identified peptides match to was determined using the NetMHCpan version 4.0 (Hoof et al., 2009), (Nielsen and Andreatta, 2016) (**Supplemental Data**). The abundance of the peptides was quantified by the MS/MS intensity values, after normalization with the summed intensity of both UC-perturbed and control cell lines. We compared the abundance and hydrophobicity (Janin, 1979) of the peptides in UCD cell lines compared to controls using Wilcoxon rank sum test.

Identifying candidate neoantigens from HLA-peptidomics data

To identify the neo-antigens, we mapped nonsynonymous mutations in UCD perturbed cells to the mass-spec data from the unperturbed and perturbed cells. The raw mass spec data was transformed to mzML format using MSConvertGUI tool, integrated in ProteoWizard 3.0 (Chambers et al., 2012). The mzML files from cell lines, each from with/without perturbation conditions, were used as an input to RAId_DbS tool (Alves et al., 2007), with all default parameters and recommended settings for our application. We allowed at most 2 missed cleavage sites. For terminal group molecular weight (Da), we chose the default 1.0078 and 17.0027 respectively for N-terminal and C-terminal attached chemical group, which accounts for the hydrogen signal and –COOH group, respectively. We used the default mass tolerance (Da) of 1.0

in precursor ion and 0.2 in product ion parameters. Finally, and used the "RAId_score" to identify peptides using P-value threshold of 0.05 (and E-value ≤ 1). We used the reference protein sequence database from NCBI (Refseq release 82) to map the peptides to protein IDs. In identifying single amino acid polymorphisms (SAPs) we allowed for all amino acids. The RAId_DbS outputs, each from the paired cell lines, were used to map the amino-acid change to non-synonymous mutations on genes, separately for R->Y and Y->R cases, reported in VCF files, using in-house python script. There are three types of peptide changes we considered in identifying neoantigens: (i) the amino acid (AA) change that is matched to mutation in dbSNP but not to R->Y mutation in the VCF mutation data of these cell lines, (ii) the AAs that are mapped to the R->Y mutation in VCF files but neither correspond to AA changes and nor are reported in dbSNP, (iii) the AA change that coincide with R->Y (or Y->R) mutation in the VCF file and is reported in dbSNP (**Table S6**). Out of the 16 neoantigens in UC perturbed cell lines 9 of them match with the cell lines VCF file, while the remainder matches to dbSNP (**Table S7**). MHC-I binding affinity was calculated based on six different methods, and percentile rank was used to determine MHC-I binding (2.0 as threshold) (Jurtz et al., 2017) (Andreatta and Nielsen, 2016), (Karosiene et al., 2012), (Zhang et al., 2009), (Rasmussen et al., 2016) (<http://www.biorxiv.org/content/early/2017/08/09/174243>). We additionally used IEDB (Vita et al., 2015) to find any additional MHC-I epitopes matching to the identified candidate neopeptides satisfying both overall >90% BLAST and 2nd/9th positions match. We compared the hydrophobicity (Janin, 1979) of the R->Y neo-antigens before and after mutation using paired Wilcoxon rank sum test.

Cell cultures

Patients fibroblast studies were performed on anonymized cells devoid of all identifiers. HepG2, SKOV3 and U2OS cell lines was purchased from ATTC. LOX-IMVI was purchased from NCI60 (RRID:CVCL1381). OTC and CPS1 deficient cell lines were purchased from Coriell Institute for Medical Research. Cells were cultured using standard procedures in a 37 °C humidified incubator with 5% CO₂ in Dulbecco's Modified Eagle's Medium (DMEM) (gibco) or RPMI 1640 medium (gibco) supplemented with 10-20% heat-inactivated fetal bovine serum, 10% pen-strep and 2 mM glutamine. MC38 cell line, derived from mouse colon adenocarcinoma (kindly provided by Dr. Eran Elinav Department of Immunology, The Weizmann Institute of Science). MCF10A and its medium was kindly provided by Professor Yossi Yarden, Department

of Biological Regulation, The Weizmann Institute of Science. All cells were tested routinely for Mycoplasma using Mycoplasma EZ-PCR test kit (#20-700-20, Biological Industries, Kibbutz Beit Ha'emek).

Crystal violet staining

Cells were seeded in 6 or 12-well plates at 50,000-200,000 cells/well in a triplicate. Time 0 was calculated as the time the cells became adherent, which was after about 10 hours from plating. For each time point, cells were washed with PBS X1 and fixed in 4% PFA (in PBS). Cells were then stained with 0.2% (for MCF10A cells)-0.5% (other cell lines) Crystal Violet (Catalog #: C0775, Sigma-Aldrich) for 20 minutes (1ml per well) and washed with water. Cells were then incubated with 10% acetic acid for 20 minutes with shaking. Extract was then diluted 1:1-1:4 in water and absorbance was measured at 595 nm 24-72 hours following time 0.

5FU and Rapamycin treatment

Survival analysis: LOX-IMVI, SKOV3, and Hepg2 (clone F10 and G03/4) perturbed cancer cells were seeded in 6-well plates at 50,000-200,000 cells/well cells per well. The following day, cells were treated with 10 μ M 5FU (F6627, Sigma-Aldrich) 5FU was renewed into the medium every day for 3 days. Survival rate of cells was quantified using Crystal violet as described above. Nucleotide synthesis: LOX-IMVI, SKOV3, and Hepg2 clone G03/4) perturbed cancer cells were seeded in 10cm plates The following day, cells were treated with 100 μ M (SKOV3) -200 μ M (LOX-IMVI and Hepg2) of Rapamycin (R0395, Sigma-Aldrich) or 10-20 μ M 5FU (F6627, Sigma-Aldrich) 5FU and Rapamycin were renewed into the medium every day for 3 days Cells treated with Rapamycin or control (DMSO Vehicle) were extracted and processed for Western blot analysis as describes below. Cells incubated with 5FU were extracted with Methanol and nucleotide synthesis levels were analysed as described below.

Western blotting

Cells were lysed in RIPA (Sigma-Aldrich) and 0.5% protease inhibitor cocktail (Calbiochem), 1% phosphatase inhibitor cocktail (P5726, Sigma-Aldrich). Following centrifugation, the supernatant was collected and protein content was evaluated by the Bradford assay or BCA Protein Assay Kit (ThermoFisher Scientific, cat # 23225) 100 μ g from each sample under reducing conditions were loaded into each lane and separated by electrophoresis on a 10% SDS polyacrylamide gel. Following electrophoresis, proteins were transferred to Cellulose Nitrate

membranes (Tamar, Jerusalem, Israel). Nonspecific binding was blocked by incubation with TBST (10 mM Tris-HCl (pH 8.0), 150 mM NaCl, 0.1 % Tween 20) containing 5% skim milk for 1h at room temperature or BSA 3% (Sigma, A7906). Membranes were subsequently incubated with antibodies against: p97 (1:10,000, PA5-22257, Thermo Scientific), GAPDH (1:1000, 14C10, #2118, Cell Signalling), CAD (1:1000, ab40800, Abcam and 1:500, Cell Signalling 11933), phospho-CAD (Ser1859) (1:1000, #12662, Cell Signalling), ASL(1:1000, ab97370, Abcam), Actin (1:1000, A5441, Sigma-Aldrich), OTC (1:1000, ab203859, Abcam), ASS1 (1:500, sc-99178 Santa Cruz), ORNT1 (1:500, NBP2-20387, Novusbio), Phospho-p70 S6 Kinase (Ser371) (1:1000 9208 Cell signalling), p70 S6 Kinase (1:500, Cell signalling 9202). Antibody was detected using peroxidase-conjugated AffiniPure goat anti-rabbit IgG or goat anti-mouse IgG (Jackson ImmunoResearch, West Grove, PA) and enhanced chemiluminescence western blotting detection reagents (EZ-Gel, Biological Industries). Gels were quantified by Gel Doc™ XR+ (BioRad) and analysed by ImageLab 5.1 software (BioRad). The relative intensity of each band was calculated by dividing the specific band intensity with the value obtained from the loading control.

Immunohistochemistry

Four micrometer paraffin embedded tissue sections were de-paraffinized and rehydrated. Endogenous peroxidase was blocked with three percent H₂O₂ in methanol. Sections undergoing for ASL, ORNT1 (SLC25A15), ASS1, OTC and PCNA staining we performed antigen retrieval in citric acid (pH 6), for 10 minutes or Tris EDTA (pH 9), using a low boiling program in the microwave to break protein cross-links and unmask antigens. After pre-incubation with 20% normal horse serum and 0.2% Triton X-100 for 1 hour at RT, biotin block via Avidin/Biotin Blocking Kit (SP-2001, Vector Laboratories, Ca, USA), sections were incubated with the primary antibodies as follow; ASL (1:50 dilution, Abcam, ab97370, CA, USA); ORNT1 (1:200 dilution, NBP2-20387, Novus Biologicals, CO, USA); ASS1 (1:50 dilution, Abcam, ab124465, CA, USA); OTC 1:100 LSBio,LS-C31865, WA, USA); PCNA (1:100. Santa Cruz, Ca, USA). All antibodies were diluted in PBS containing 2% normal horse serum and 0.2% Triton. Sections were incubated overnight at RT followed by 48h at 4°C. Sections were washed three times in PBS and incubated with secondary biotinylated IgG at RT for 1.5 hour, washed three times in PBS and incubated with avidin-biotin Complex (Elite-ABC kit, Vector Lab) at RT for additional 90 min followed by DAB (Sigma) reaction. Stained sections were examined and photographed

by a bright field microscope (E600, Tokyo, Japan) equipped with Plan Fluor objectives (10x) connected to a CCD camera (DS-Fi2, Nikon). Digital images were collected and analyzed using Image Pro+ software. Images were assembled using Adobe Photoshop (Adobe Systems, San Jose, CA).

Virus infection

Primary fibroblasts were infected with HCMV and harvested at different times after infection for ribosome footprints (deep sequencing of ribosome-protected mRNA fragments) as previously described (Tirosh et al., 2015). Briefly we infected human foreskin fibroblasts (HFF) with the Merlin HCMV strain and harvested cells at 5, 12, 24 and 72 hours post infection. Cells were pre-treated with Cycloheximide and ribosome protected fragments were then generated and sequenced. Bowtie v0.12.7 (allowing up to 2 mismatches) was used to perform the alignments. Reads with unique alignments were used to compute footprints densities in units of reads per kilobase per million (RPKM).

Cancer cells or MCF10A cells (kindly provided by Yossi Yarden lab, Department of Biological Regulation, The Weizmann Institute of Science) were infected with either pLKO-based lentiviral vector with or without the human OTC and SLC25A15, ASS1, GFP short hairpin RNA (shRNA) (Dharmacon). Infected cells were selected with 2-4 $\mu\text{g ml}^{-1}$ puromycin.

RNA processing and quantitative PCR.

RNA was extracted from cells by using RNeasy Mini Kit (Qiagen # 74104. cDNA was synthesized from 1 μg RNA by using qScript cDNA Synthesis Kit (Quanta #95749).

Detection on cDNAs was performed using either SYBR green PCR master mix (Thermo Fisher scientific #4385612) or TaqMan Fast Advanced Master Mix (ThermoFisher scientific #4444557), with the required primers. Primer sequences are summarized in star methods key resources table.

Transient transfection

LOX-IMVI melanoma cells were seeded in 6-well plates at 70,000cells/ well, or in 12-well plates at 100,000cells/ plate. At the following day, cells were transfected with either 700pmol or 350pmol siRNA siGenome SMARTpool targeted to human SLC25A13 mRNA (#M-007472-01, Dharmacon), respectively.

Hepatocellular and ovarian carcinoma cells were seeded in 6-well plate at 10^6 or 70,000 cells/well respectively, transfected with 2-3 μ g of the OTC (EXa3688-LV207 GENECOPOEIA) or ORNT1 (EXu0560-LV207 GENECOPOEIA) plasmids. Transfection was done with Lipofectamine® 2000 Reagent (#11668027, ThermoFisher Scientific), in the presence of Opti-MEM® I Reduced Serum Medium (#11058021, ThermoFisher Scientific). 4 hours after transfection, medium was replaced and the experiments were performed 48-108 hours post transfection.

Overexpression

LOX-IMVI melanoma cells were transduced with pLEX307-based lenti viral vector with or without the human SLC25A13 transcript, encoding for citrin. Transduced cells were selected with 2 μ g/ml Puromycin.

Metabolomics analysis

Urea and uracil: Each cell line was seeded at $3 - 5 \times 10^6$ cells per 10 cm plate and when confluent, washed with ice-cold saline, lysed with a mixture of 50% methanol in water added with 2 μ g/mL ribitol as an internal standard and quickly scraped followed by three freeze-thaw cycles in liquid nitrogen. The insoluble material was pelleted in a cooled centrifuge (4 °C) and the supernatant was collected for consequent GC-MS analysis. Samples were dried under air flow at 42 °C using a Techne Dry-Block Heater with sample concentrator (Bibby Scientific) and the dried samples were treated with 40 μ l of a methoxyamine hydrochloride solution (20 mg ml⁻¹ in pyridine) at 37 °C for 90 min while shaking followed by incubation with 70 μ l N,O-bis(trimethylsilyl) trifluoroacetamide (Sigma) at 37 °C for an additional 30 min.

Isotopic labelling: Hepatocellular and ovarian carcinoma cells were seeded in 10 cm plates and once the cell confluence reached 80%, cells were incubated with 4mM L-GLUTAMINE, (ALPHA-15N, 98%, Cambridge Isotope Laboratories, Inc.) for 24 hours. Subsequently, cells were processed as described above. GC-MS analysis used a gas chromatograph (7820AN, Agilent Technologies) interfaced with a mass spectrometer (5975 Agilent Technologies). An HP-5ms capillary column 30 m \times 250 μ m \times 0.25 μ m (19091S-433, Agilent Technologies) was used. Helium carrier gas was maintained at a constant flow rate of 1.0 ml min⁻¹. The GC column temperature was programmed from 70 to 150 °C via a ramp of 4 °C min⁻¹, 250-215 °C via a ramp of 9 °C min⁻¹, 215-300 °C via a ramp of 25 °C min⁻¹ and maintained at 300 °C for an additional 5 min. The MS was by electron impact ionization and operated in full-scan mode from

$m/z = 30-500$. The inlet and MS transfer line temperatures were maintained at 280 °C, and the ion source temperature was 250 °C. Sample injection (1 -3 μl) was in splitless mode.

Nucleotides analysis

Materials: Ammonium acetate (Fisher Scientific) and ammonium bicarbonate (Fluka) of LC-MS grade were used. Sodium salts of AMP, CMP, GMP, TMP and UMP were obtained from Sigma-Aldrich. Acetonitrile of LC grade was supplied from Merck. Water with resistivity 18.2 M Ω was obtained using Direct 3-Q UV system (Millipore).

Extract preparation: The obtained samples were concentrated in speedvac to eliminate methanol, and then lyophilized to dryness, re-suspended in 200 μl of water and purified on polymeric weak anion columns (Strata-XL-AW 100 μm (30 mg ml⁻¹, Phenomenex)) as follows. Each column was conditioned by passing 1 ml of methanol, then 1 ml of formic acid/methanol/water (2/25/73) and equilibrated with 1 ml of water. The samples were loaded, and each column was washed with 1 ml of water and 1 ml of 50% methanol. The purified samples were eluted with 1 ml of ammonia/methanol/water (2/25/73) followed by 1 ml of ammonia/methanol/water (2/50/50) and then collected, concentrated in speedvac to remove methanol and lyophilized. Before LC-MS analysis, the obtained residues were re-dissolved in 100 μl of water and centrifuged for 5 min at 21,000 g to remove insoluble material.

LC-MS analysis: The LC-MS/MS instrument consisted of an Acquity I-class UPLC system (Waters) and Xevo TQ-S triple quadrupole mass spectrometer (Waters) equipped with an electrospray ion source and operated in positive ion mode was used for analysis of nucleoside monophosphates. MassLynx and TargetLynx software (version 4.1, Waters) were applied for the acquisition and analysis of data. Chromatographic separation was done on a 100 mm \times 2.1 mm internal diameter, 1.8- μm UPLC HSS T3 column equipped with 50 mm \times 2.1 mm internal diameter, 1.8- μm UPLC HSS T3 pre-column (both Waters Acquity) with mobile phases A (10 mM ammonium acetate and 5 mM ammonium hydrocarbonate buffer, pH 7.65 adjusted with 10% acetic acid) and B (acetonitrile) at a flow rate of 0.3 ml min⁻¹ and column temperature 25 °C. A gradient was used as follows: for 0-3 min the column was held at 0.2% B, then 3-3.5 min a linear increase to 100% B, 3.5-4.0 min held at 100% B, 4.0-4.5 min back to 0.2% B and equilibration at 0.2% B for 2.5 min. Samples kept at 8 °C were automatically injected in a volume of 3 μl . For mass spectrometry, argon was used as the collision gas with a flow of 0.10

ml min⁻¹. The capillary voltage was set to 2.50 kV, source temperature 150 °C, desolvation temperature 400°C, cone gas flow 150 l hr⁻¹, desolvation gas flow 800 l hr⁻¹.

Nucleotide concentration was calculated using a standard curve of the relevant nucleotide concentration in each sample. Standard curves included increasing concentration of all measured nucleotides ranging from 0-10ug/ml that were positioned at the beginning and at the end of each run. All the calculated values for the different nucleotides in each sample fell within the standard curve range. Analytics were detected in positive mode using multiple-reaction monitoring listed in supplementary table 14.

Patient samples

All prostate patients' urine samples were obtained upon informed consent and with evaluation and approval from the corresponding ethics committee (CEIC code OHEUN11-12 and OHEUN14-14(Royo et al., 2016)). Patients included in the study were diagnosed with prostate adenocarcinoma and the criteria for inclusion was to be scheduled for surgery as anticancer treatment. Samples were collected between 2012 and 2016.

Urea analysis: To estimate age-stratified Urea background levels we pooled data from X=1,363,691 patients in the de-identified Clalit Health Services electronic health record (code 0194-17-COM2), and y=100 pediatric cancer patients on their day of admission to the Pediatric Hemato-Oncology Department at Souraski Medical Center (code 0016-17). Median urea level was computed per sample per year and values were grouped by age and the distribution was summarized in a box plot. Cases were analyzed in a similar fashion. P-values were estimated using a MW test that was performed following additional stratification to gender.

Extraction of polar metabolites from Urine and Plasma

To extract polar metabolites from urine (20~100 uL) and plasma (100 uL) samples, 1 and 0.9 mL methanol (with labeled amino acids as internal standard) were added, respectively, into biological sample-containing Eppendorf tube. Then, the resulting mixture was vortexed and sonicated for 15 min, vortexed again and centrifuged at 14000 rpm for 10 min. The liquid phase was transferred into new tube and lyophilized. Then the pellets were dissolved using 150 uL DDW-methanol (1:1), centrifuged twice to remove possible precipitants, and was injected into LC-MS system.

LC-MS polar metabolites analysis

Metabolic profiling of polar phase was done as described at Liang Zheng et al. (2015) with minor modifications described below. Briefly, analysis was performed using Acquity I class UPLC System combined with mass spectrometer (Thermo Exactive Plus Orbitrap) which was operated in a negative ionization mode. The LC separation was done using the SeQuant Zic-pHilic (150 mm × 2.1 mm) with the SeQuant guard column (20 mm × 2.1 mm) (Merck). The Mobile phase A: acetonitrile and Mobile phase B: 20 mM ammonium carbonate plus 0.1% ammonia hydroxide in water. The flow rate was kept at 200 $\mu\text{l min}^{-1}$ and gradient as follow: 0 - 2min 75% of B, 17 min 12.5% of B, 17.1 min 25% of B, 19 min 25% of B, 19.1 min 75% of B, 19 min 75% of B.

Polar metabolites data analysis

The data processing was done using TraceFinder Thermo Fisher software were detected compounds were identified by Retention time and fragments and verified using in-house mass spectra library. Urine metabolites were normalized by creatinine peak area.

In vivo experiments

Animal experiments were approved by the Weizmann Institute Animal Care and Use Committee Following US National Institute of Health, European Commission and the Israeli guidelines (IACUC 21131015-4). 8 weeks old C57BL/6 and BALB/c male and female mice were respectively injected with CT26 colon cancer cells (sub-cutaneous), or with 4T1 breast cancer cells (in the mammary pad). Three weeks following the injection an advanced tumor was observed and palpated. Control urine was obtained from C57BL/6 and BALB/c mice similar in age which were not injected. Samples below 100ul were excluded from the analysis. Urine was collected from mice presenting adverse tumors and analysed for metabolites at Baylor College of medicine or at the Life Sciences Core Facilities, Weizmann Institute of Science After sacrifice tumors and livers were removed from the mice for further analysis by quantitative PCR and Western blot.

Xenograft mouse model: 8 weeks old male SCID mice were injected sub-cutaneous with 1×10^6 cells of Hepg2 EV , Hepg2, shOTC SKOV EV and SKOV shORNT in 200ul of DMEM (Hepg2) or RPMI (SKOV) containing 5% matrigel. Tumor weight was recorded post scarification.

Syngeneic mouse models: 8 weeks old C57BL/6 male mice were injected sub-cutaneous in the right flank with MC38 mouse colon cancer cells infected with either an empty vector or with

shASS1. For each injection, 5×10^5 tumor cells were suspended in 200ul DMEM containing 5% matrigel. Following injection, mice were treated with 250ug of anti PD-1 antibody on days on days 7, 12, 16, 19 (Clones 29F.1A12 ,RPM114, Bio Cell).

On day 22, mice were euthanized and tumors were removed and incubated in 1ml of PBS containing Ca^{2+} , Mg^{2+} (Sigma D8662) with 2.5 mg/ml Collagenase D (Roche) and 1mg/ml DNase I (Roche). Following 20min incubation at 37c, the tumors were processed into a single cell suspension by mechanically grinding on top of wire mesh and repeated washing and filtering onto 70 μM filter (Falcon). Single cell suspensions from tumors were stained for flow cytometry analysis with CD3-FITC (clone 17A2), CD4-PE (clone GK15) and CD8a-APC (clone 53-6.7) all from Biolegend. Next, the cells were fixed using BD cytofix/cytoperm solution (BD Biosciences) and acquired on LSRII flow cytometer at the Weizmann FACS facility and analyzed with FlowJo software (Tree Star).

The tumor volume was quantified by the formula, $(l \times w \times h) \pi/6$, and normalized by their volume on day 11 when the mean tumor volume reached around 100 mm^3 . The response to anti-PD1 therapy (and empty vector) was quantified by the tumor volume change at time t , $\Delta V_t = (V_t - V_0)/V_0$, where V_t denotes the normalized tumor volume at a given time t , and V_0 denotes the tumor volume on day 11. The overall response of treated and control groups was compared by Wilcoxon ranksum test of ΔV_t on day 21, and the sequential tumor growth was compared using ANOVA over the whole period (where the internal tumor volume was measured on day 9, 13,17, and 19).

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical an analysis

Unless otherwise specified, all statistical analyses were performed using one-way ANOVA, Student's t-test or Wilcoxon rank sum test of multiple or two groups, with Dunnett's correction when required. Log-transformed data were used where differences in variance were significant and variances were correlated with means. The sample size was chosen in advance based on common practice of the described experiment and is mentioned for each experiment. Each experiment was conducted with biological and technical replicates and repeated at least three times unless specified otherwise. When samples were distributed non-normally, Mann-Whitney analysis was performed. Statistical tests were done using Statsoft's STATISTICA, ver. 10. All

error bars represent SER. $P < 0.05$ was considered significant in all analyses (* denotes $P < 0.05$, ** $P < 0.005$, *** $P < 0.0005$, **** $P < 0.0001$).

DATA AND SOFTWARE AVAILABILITY

The custom codes used to generate one of the core results of the manuscript are available in github repository: <https://www.github.org/jooslee/UCD>, and the full custom codes will be deposited. The raw sequencing data from the cell line experiments is available through European Nucleotide Archive (<https://www.ebi.ac.uk/ena>, ID: PRJEB27646), and processed data is available through Mendeley Data at doi:10.17632/xc4pj5py86.1.

ADDITIONAL RESOURCES

N/A

Supplemental Tables

Table S1. Related to Figure 1. Fraction of the samples of UC dysregulated and PTMB in different cancer types

Table S2. Related to Figure 3. Increased nitrogen utilization for pyrimidine synthesis following UCD

Table S3. Related to Figure 4. Pyrimidine-rich mutational bias of UC-perturbed cancer cell lines

Table S4. Related to Figure 5. Association of PTMB with worse prognosis

Table S5. Related to Figure 6. Accuracy of UCD and PTMB compared to that of mutational load in predicting ICT response

Table S6. Related to Figure 6. Identities of neo-antigens in UC-perturbed cancer cell lines

Table S7. Related to Figure 6. Neo-antigen bias of UC-perturbed cancer cell lines

Table S8. Related to Figure 6. The codon frequency table of the four SNP types (RY, RR, YY, YR)