Genomic evolution shapes prostate cancer disease type

Dan J. Woodcock[1,2,3], Atef Sahli[2,3], Ruxandra Teslo[3], Vinayak Bhandari[4], Andreas J. Gruber[2,3,5], Aleksandra Ziubroniewicz[1,3], Gunes Gundem[6,7], Yaobo Xu[6], Adam Butler[6], Ezequiel Anokian[8], Bernard J. Pope[9,10,11], Chol-Hee Jung[9], Maxime Tarabichi[12,13], Stefan C. Dentro[2,6,12], J. Henry R. Farmery[14], CRUK ICGC Prostate Group, Peter Van Loo[12,15,16], Anne Y. Warren[17], Vincent Gnanapragasam[18,19,20], Freddie C. Hamdy[1], G. Steven Bova[21,22], Christopher S. Foster[23], David E. Neal[24,25], Yong-Jie Lu[26], Zsofia Kote-Jarai[8], Michael Fraser[4], Robert G. Bristow[4,27,28,29,30], Paul C. Boutros[4,31,32], Anthony J. Costello[33,34], Niall M. Corcoran[33,34], Christopher M. Hovens[33,34], Charlie E. Massie[24,35,**], Andy G. Lynch[14,36], Daniel S. Brewer[37,38,43,*], Rosalind A. Eeles[8,39,43,*], Colin S. Cooper[8,37,43,*], David C. Wedge[2,3,30,40,41,43,44,*]

[1] *Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK*
[2] *Nuffield Department of Medicine, University of Oxford, Oxford, UK*
[3] *Big Data Institute, University of Oxford, Oxford, UK*
[4] *Department of Medical Biophysics, University of Toronto, Toronto, Canada*
[5] *Department of Biology, University of Konstanz, Konstanz, Germany*
[6] *Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK*
[7] *Memorial Sloan-Kettering Cancer Center, New York, USA*
[8] *The Institute of Cancer Research, London, UK*
[9] *Melbourne Bioinformatics, University of Melbourne, Melbourne, Victoria, Australia*
[10] *Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia*
[11] *Department of Medicine, Central Clinical School, Monash University, Melbourne, Victoria, Australia*
[12] *The Francis Crick Institute, London, UK*
[13] *Institute of Interdisciplinary Research (IRIBHM), Universite Libre de Bruxelles, Brussels, Belgium*
[14] *Statistics and Computational Biology Laboratory, Cancer Research UK Cambridge Institute, Cambridge, UK*
[15] *Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[16] *Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*
[17] *Department of Histopathology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK*
[18] *Cambridge Urology Translational Research and Clinical Trials Office, Addenbrooke's Hospital, Cambridge, UK*
[19] *Division of Urology, Department of Surgery, University of Cambridge, Cambridge, UK*
[20] *Department of Urology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK*
[21] *Prostate Cancer Research Center, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland*
[22] *Tays Cancer Center, Tampere University Hospital, Tampere, Finland*
[23] *HCA Laboratories, London, UK*
[24] *Uro-Oncology Research Group, Cancer Research UK Cambridge Institute, Cambridge, UK*
[25] *Department of Surgical Oncology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK*
[26] *Centre for Molecular Oncology, Barts Cancer Institute, Queen Mary University of London, London, UK*

[27] Division of Cancer Sciences, Faculty of Biology, Health and Medicine, University of Manchester, Manchester, UK
[28] The Christie NHS Foundation Trust, Manchester, UK
[29] CRUK Manchester Institute, University of Manchester, Manchester, UK
[30] Manchester Cancer Research Centre, University of Manchester, Manchester, UK
[31] Departments of Human Genetics and Urology, University of California, Los Angeles, Los Angeles, California, USA
[32] Jonsson Comprehensive Cancer Center, University of California, Los Angeles, Los Angeles, California, USA
[33] Departments of Urology and Surgery, Royal Melbourne Hospital and The University of Melbourne, Melbourne, Victoria, Australia
[34] Victorian Comprehensive Cancer Centre, Parkville, Victoria, Australia
[35] Early Detection Programme and Urological Malignancies Programme, Cancer Research UK Cambridge Centre, Department of Oncology, University of Cambridge, Cambridge, UK
[36] School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, UK
[37] Norwich Medical School, University of East Anglia, Norwich, UK
[38] Earlham Institute, Norwich, UK
[39] Royal Marsden NHS Foundation Trust, London, UK
[40] Oxford NIHR Biomedical Research Centre, Oxford, UK
[41] Manchester NIHR Biomedical Research Centre, Manchester, UK
[42] A list of members and affiliations appears in the STAR Methods
[43] Senior authors
[44] Lead contact

Abstract

The development of cancer is an evolutionary process involving the sequential acquisition of genetic alterations that disrupt normal biological processes, enabling tumour cells to rapidly proliferate and eventually invade and metastasise to other tissues. We investigated the genomic evolution of prostate cancer through the application of three separate classification methods, each designed to investigate a different aspect of tumour evolution. Integrating the results revealed the existence of two distinct types of prostate cancer that arise from divergent evolutionary trajectories, designated as the Canonical and Alternative evolutionary disease types. We therefore propose the evotype model for prostate cancer evolution wherein Alternative-evotype tumours diverge from those of the Canonical-evotype through the stochastic accumulation of genetic alterations associated with disruptions to androgen receptor DNA binding. Our model unifies many previous molecular observations, providing a powerful new framework to investigate prostate cancer disease progression.

[*] Corresponding authors: d.brewer@uea.ac.uk, colin.cooper@uea.ac.uk, ros.eeles@icr.ac.uk, david.wedge@manchester.ac.uk
[**] Deceased

1. Introduction

Tumour evolution is a dynamic process[1] involving the accumulation of genetic alterations that disrupt normal cellular processes, leading to pathological phenotypes[2]. While some cancers can be categorised into subtypes, often utilising pronounced genomic or transcriptomic differences, the evolutionary processes that give rise to this variation are complex and not well understood[3]. However, it has been shown that the order of events in some haematological malignancies can be related to prognosis and treatment susceptibility[4,5,6,7].

In prostate cancer, subtyping schemes have been proposed based on the presence of specific molecular alterations[8], combinations of alterations[9] or gene expression profiles[10]. However, detailed investigations by ourselves[11] and others[12,13] have shown substantial heterogeneity between tumours that presents challenges for simple or consistent subtype assignments[14]. Studies investigating evolutionary differences between prostate cancer disease types by categorising molecular events as "early" or "late" have been shown to be informative in early-onset[15] and aggressive disease[16] and the temporal order of genetic alterations has also been shown to be related to the ETS subtype[11]. However, the evolutionary factors that drive the emergence of prostate cancer subtypes remains largely unexplored.

To investigate how evolutionary behaviour manifests in the variation observed in prostate cancer genomes, we performed three separate analyses, each of which probes different aspects of tumour evolution. In each analysis we classified the tumours in an unsupervised fashion and subsequently identified sets of tumours that shared the same classes across the analyses. Through this approach we can identify tumours that display consistent evolutionary properties and use this information to identify likely mechanisms driving prostate cancer evolution.

2. Results

Data collection and pre-processing

We compiled a data set from 159 intermediate or low risk prostate adenocarcinoma patients sampled after radical prostatectomy, which were otherwise treatment naïve (87 published previously[11]). These were whole genome sequenced (target depth: 50X) along with matched blood controls (target depth: 40X), and 123 summary measurements were generated (STAR Methods; Figure S1).

We adapted an unsupervised neural network with a single hidden layer to perform feature learning on this data set, identifying associations between inputs to obtain a reduced-dimension set of 30 *features* (STAR Methods). Using the trained neural network, we can recast the data for each sample in terms of these features into a form known as the *feature representation.* Reconstructing the original inputs from the feature representation gave a reconstruction error of $\approx$12%, indicating that these features, and the inputs to which they correspond, contain a substantial proportion of the information in the original data. Our approach is a white-box method meaning we can identify which inputs contribute to each feature, and so we labelled each features with a brief descriptor of the associated genomic aberrations (Figure S2). We can perform analysis on the feature representation itself, while allowing comparison with the results of other analyses using a selection of the original inputs that correspond to the features (STAR Methods).

Classifying tumours by patterns of co-occurring genomic features

Despite the reduced dimensionality of the feature representation, application of standard clustering methods remains problematic due to the high dimension of features (30) relative to the sample size (159). To mitigate this, we adopted a two-stage clustering method utilising a discrimination score we calculated for each feature that quantified the value of each feature in predicting disease relapse (STAR Methods). In the first stage, we applied k-medoid clustering to the feature representation of those features with a high discrimination score (STAR Methods). In the second stage we performed hierarchical clustering on the cluster centres (medoids) returned in the first stage. The results are shown in Figure 1.

We identified two distinct *metaclusters* that were characterised by different sets of aberrations. Metacluster A (MC-A) showed a high probability of features corresponding to intra-chromosomal structural variants (SVs), *SPOP* mutations, chromothripsis, and loss of heterozygosity (LOH) in regions 5q15-5q23.1 (spanning *CHD1*) and 6q14.1-6q22.32 (*MAP3K7*, *ZNF292*). Metacluster B showed more frequent ETS fusions, as well as LOH affecting 17p (*TP53*) and regions 19p13.3-13.2 and 22q11.21-22q11.22. The dendrogram indicated additional differences within Metacluster B and so we further divided it into subclasses, MC-B1 and MC-B2, with MC-B2 displaying near-ubiquitous *TP53* LOH and exhibiting higher probability of ETS fusions, inter-chromosomal chained structural variants (cSVs), LOH at 10q23.1-10q25.1 (*PTEN*) and 5q11.1-5q14.1 (*IL6ST*, *PDE4D*).

Classifying tumours by mechanism of DNA double strand breaks.

We investigated the influence of Androgen Receptor (AR) on the DNA breakpoints in these samples. AR is known to precipitate DNA double strand breaks (DSB) in conjunction with topoisomerase II-beta[17], and AR-associated breakpoints are frequent in early-onset prostate cancer[15,18]. Furthermore, it has also been shown that AR binding behaviour can be altered by *CHD1* deletion[19]. We used a permutation test (STAR Methods) to classify tumours based on whether breakpoints occurred significantly more (labelled as Enriched) or less (Depleted) often proximal to AR binding sites (ARBS) than expected if they were independent of AR, or Indeterminate tumours that displayed no statistically significant association (Figure 2A).

Investigating the ARBS groups in conjunction with the genetic alterations associated with the features (Figure 2B), we found that Depleted tumours had the highest percentage genome altered (PGA) and the highest frequency of multiple CNAs, chromothripsis, kataegis, and *SPOP* mutations (Relationship column, Figure 2B). Enriched and indeterminate tumours displayed no significant differences for any CNAs, but both showed higher frequency of CNAs covering *PTEN* and *TP53* than the Depleted group (Relationship column, Figure 2B). In the case of ETS fusions and inter/intra-chromosomal cSV ratio, the Enriched group showed greater amounts than the intermediate group, which in turn showed greater enrichment than the Depleted group. Both Enriched and Depleted tumours displayed higher numbers of breakpoints than Indeterminate tumours. We identified these ARBS groups in two additional data sets: a set of low-intermediate risk tumours from the Canadian Prostate Cancer Genome Network (CPC-GENE)[13], and high-risk tumours from the Melbourne Prostate Cancer Research Group in Australia (unpublished). Clustering these groups by CNA proportions showed groups classified as Depleted clustered together (Figure S3), confirming the association between these CNAs and ARBS-distal breakpoint prevalence.

Classifying tumours through the evolutionary order of key events

The order in which genetic alterations generally occur in tumour evolution, subsequently referred to as the 'ordering profile', can be inferred using the estimated proportion of tumour cells that display each genetic alteration in each sample[11]. We adapted a Plackett-Luce mixture model[20] to create a probabilistic model for the relative order of genomic aberrations given the relative subclonal fractions of *SPOP* mutations and the key CNAs that were identified in our feature extraction (STAR Methods). As a mixture model, it can be used to extract distinct ordering profiles within the population. Inference with this model was performed with differing numbers of clusters, and the results used in Bayesian model selection that determined that two ordering profiles was optimal (STAR Methods). We therefore defined two classes, Ordering-I and Ordering-II, and each tumour was assigned to one of these by their mixture weights (Figure 3).

The two profiles displayed notable differences. Tumours corresponding to Ordering-I frequently experienced an early 8p LOH (spanning *NKX3.1*) and ETS fusions. Less frequent LOH of regions covering the *RB1*, *BRCA2*, *CDH1*, *TP53* or *PTEN* genes could also occur. This profile occasionally displayed a very early LOH of 1q42.12-42.3. Tumours of Ordering-II consistently displayed early LOH events covering *MAP3K7* and 13q (*EDNRB*, *RB1*, *BRCA2*). However, the earliest events, a mutation of the *SPOP* gene and LOH covering *CHD1*, were less frequent. Ordering-II also displayed more frequent copy number gains. Both orderings showed late gains of chromosome 19. When comparing the occurrence of aberrations between individuals within each Ordering we found that the relative order of alterations was highly variable, indicating they arise stochastically (Figure S4; STAR Methods).

Integrating analyses reveals disease types distinguished by their evolutionary trajectories.

Establishing the concordance of these three classification methods (Figure 4A) revealed a remarkable relationship: MC-A is largely a subset of the Depleted group (22/27), and both are almost entirely subsets of Ordering-II (26/27 and 30/32 respectively). Quantifying the strength of the pairwise associations using Cramer's V statistic gives: Metaclusters and ARBS groups ($V = 0.69$), Metaclusters and Orderings ($V = 0.58$) and ARBS and Orderings ($V = 0.62$). These values indicate a strong association between cluster assignments in these three groups (STAR methods). We can therefore infer that there exists a subset of tumours that exhibit all the corresponding properties: an evolutionary trajectory (Ordering-II), a breakpoint mechanism (ARBS:Depleted) and characteristic patterns of aberrations (Metacluster:MC-A). We therefore propose the *evotype model for prostate cancer evolution* (Figure 4B), in which canonical AR DNA binding is disrupted, through the effect of genetic alterations or other causes, coercing tumour evolution along an alternative trajectory that results in a distinct form of the disease. We can therefore classify tumours by which path a tumour is most likely to adhere to, which we refer to as its "evotype". To perform this classification, we adopted a majority-vote approach and defined tumours that were assigned to at least two of MC-A, Depleted, or Ordering-II as belonging to the *Alternative-evotype* (n=34), to distinguish them from *Canonical-evotype* (n=125) tumours that evolve via the standard route. Each evotype is characterised by a different propensity for certain aberrations (Figure 4C), but we found that no single aberration was either necessary or sufficient for assignment to either evotype. However, there were several pairwise combinations of genetic alterations that did result in fixation to one of the evotypes (Figure S5). There were no statistically significant associations (p =0.05) between the evotypes and tumour stage, Gleason grade or prostate-specific antigen (PSA) levels (Figure S6).

The lack of consistent genetic alterations indicates that there may be multiple individual routes of progression for each evotype. We investigated these trajectories in more detail by developing a stochastic model of the acquisition of genetic alterations and tracking the probability of assignment to each evotype as the aberrations accumulate (Figure 4D; STAR Methods). Initially the probability density is concentrated at ≈0.78, the proportion of Canonical-evotype tumours in our sample set. As the number of aberrations increases, the density diverges to accumulate at 1 (corresponding to unambiguous assignment to the Canonical-evotype) and 0 (Alternative-evotype). In this model, an individual tumour will follow a trajectory through this probability landscape dependent on the type and order of aberrations. Due to randomness in the occurrence of genetic alterations, there are an enormous number of possible routes, but investigating patterns of aberrations in areas of high probability density reveals common modes of behaviour (Figures S7 and S8). Exemplars for these modes are given by the dashed lines in Figure 4D. Notably, when an *SPOP* mutation occurs first, it confers high probability (≈0.91) of progression to the Alternative-evotype (*Alternative:Rapid*). Other routes to the Alternative-evotype involve the accumulation of multiple individual LOH events involving genes such as *MAP3K7*, *CHD1* or *EDNRB* (*Alternative:Incremental*) in any order. LOH of *IL6ST* or gain of region 8p23.3-8p22 strongly influence convergence after a number of aberrations have already accumulated (*Alternative:Abrupt*). Conversely, fixation to the Canonical-evotype is dependent on a few key aberrations. Early *TP53* loss or *ERG* gene fusion promotes almost certain fixation to the Canonical-evotype (*Canonical:Rapid*). Alternatively, loss of regions covering *PTEN* or *CDH1* can coerce a relatively quick progression toward this evotype, but these are rarely the final convergent event in the trajectory (*Canonical:Moderate*). Indeed, there are aberrations that are often the last step in convergence to the Canonical-evotype, particularly LOH of 19p13.3-19p13.2 or 22q11.21-22q11.22, or gains of chromosome 19 or region 22q11.1-22q11.23 (*Canonical:Punctuated*).

The lack of a single genetic alteration unique to the Alternative-evotype indicates that there may be multiple mechanisms for acquired AR dysregulation that we observe in prostate cancer. We therefore investigated potential mechanisms of AR dysregulation. It has previously been shown that *CHD1* protein is involved in AR binding, which causes DNA loops that can precipitate DSBs (Figure 5A, adapted from Metzger et al.[21]). As LOH of the *CHD1* locus is significantly associated with the Alternative-evotype (Figure 4C) and is an early event in tumour evolution (Figure 3), we hypothesised that loss of *CHD1* in these tumours would be associated with fewer DSBs precipitated through the DNA loop mechanism. We therefore tested whether pairs of adjacent AR binding sites required for DNA loops to form by this mechanism are significantly more or less frequent close to DSBs dependent on *CHD1* status (STAR methods). We found that *CHD1*wt tumours more frequently displayed DSBs close to pairs of AR binding sites than tumours that displayed a *CHD1*-associated LOH (Figure 5B, $p = 0.00025$). Extrapolating our hypothesis to the evotypes, we found a significant difference between Canonical and Alternative-evotype tumours (Figure 5C, ($p = 4.91 \text{ x } 10^{-9}$). This relationship also holds in *CHD1*wt tumours of both evotypes (Figure 5D, $p = 0.00015$). These results indicate that *CHD1* LOH can drive AR-dysregulation in prostate cancer, but that other mechanisms also exist in Alternative-evotype tumours.

3. Discussion

Taken together, our findings reveal prostate cancer disease types that arise as a result of divergent trajectories of a stochastic evolutionary process in which specific genetic alterations can tip the balance toward convergence to either route. Unlike the evolution of species, which involves ongoing adaptation to a perpetually changing environment, tumour evolution has a definable end point - a disease state that leads to the death of the host. It follows that the more "evolved" tumours are closer to this end point, which has obvious implications for risk stratification. We therefore proposed that our evolutionary model implied two factors associated with risk, the evotype itself and the degree of progression relative to that evotype. We investigated this principle using follow-up information based on time to biochemical recurrence (serum PSA $>$ 0.2 ng/ml for two consecutive measurements) after prostatectomy.

Initially, we found that classifying by evotype alone provides a significant association with time to biochemical recurrence (Figure 6A, p = 0.026), displaying a higher hazard ratio (HR = 2.30) than stratification by well-known genetic alterations such as *PTEN* loss (HR = 1.42, p = 0.336; Figure S9A), *TP53* loss (HR = 2.03, p =0.0497; Figure S9B) or ETS status (HR = 1.64, p = 0.179; Figure S9C). However, it performed worse than other metrics known to be associated with outcome, such as tumour mutational burden (TMB), which led to HR = 4.50 and p = 0.000110 (Figure 6B), or histopathological grading via the ISUP Gleason grade score, which gave HR= 4.69 and p =0.0000629 (Figure 6C).

To illustrate how information on the evolutionary path might improve risk stratification, we adopted two approaches to determining which tumours were the most advanced relative to their evotype. In the first, we classified the 10 tumours of both evotypes with the highest TMB as advanced (denoted High-TMB Alternative and High-TMB Canonical), and compared these to all other tumours. We found that High-TMB tumours of both evotypes displayed high hazard ratios (Figure 6D; HR $>$ 6) compared to all previous metrics, notably outperforming the 20 High-TMB tumours when evotype was not used (Figure 6B; HR = 4.50). To investigate how this risk determinant might be used in conjunction with current clinical prognostic methods, we compared the 10 High-TMB tumours of both evotypes that were also ISUP Gleason grade $\geq$ 3 with all other tumours, which further improved performance (HR = 7.28, p = 5.16 x 10^{-7}; Figure 6E). In the second approach, we hypothesised that metaclusters MC-A and MC-B2 were representative of advanced tumours of the Alternative and Canonical-evotypes respectively, as these tumours displayed many of their characteristic genetic alterations (Figure 1). Stratifying by tumours belonging to both MC-A and the Alternative-evotype yielded HR = 3.64, p =0.00363, with those in MC-B2 and the Canonical-evotype giving HR = 6.14, p = 4.60 x 10^{-5}, in comparison to the tumours that were in neither group (Figure 6F). These relationships were still significant when adjusted for TMB, Gleason grade, and age at diagnosis (p_{adj} =0.00913 and 0.000492), showing that TMB itself is not driving this result. As before, we compared these advanced tumours that were also ISUP Gleason grade $\geq$ 3 to all other tumours, which provided even better performance (HR = 7.66, p = 2.84 x 10^{-8}; Figure 6G). The findings in Figures 6E and 6G indicate that Gleason grade and evolutionary progression provide complementary information on prognosis. Note that these findings are illustrative as a robust optimisation of thresholds or sets of genetic alterations for risk evaluation requires full validation with an independent data set and therefore remains outside the scope of this study.

Furthermore, the evotype model provides additional context to relationships between individual aberrations reported in previous studies. Co-occurring genomic alterations that have been identified previously can be related to particular evotypes. For the Canonical-evotype, this includes LOH events affecting *PTEN* and CDH[22], or *PTEN* and *TP53*[23]. Conversely, *CHD1* losses have previously been observed in conjunction with *SPOP* mutations[24,25], as has LOH affecting *MAP3K7*[26] and 2q22[27]; all these aberrations are associated with the Alternative-evotype. The most widely used basis for genomic prostate cancer subtyping is the ETS status, where tumours are classified by the presence or absence of an ETS gene fusion into ETS+ and ETS- respectively[8,9,1,11]. We found that 94% of Alternative-evotype tumours were ETS-, and indeed alterations such as *SPOP* mutations and *CHD1* LOH that are characteristic of this evotype have previously been associated with ETS- tumours[11,28]. Conversely, the Canonical-evotype exhibits both ETS+ (66%) and ETS- (34%) tumours. When removing Alternative-evotype tumours from the ETS classification, we found that there were no significant differences in risk (Figure S9D) or prevalence of any of the genomic features between ETS+ and ETS- tumours of the Canonical-evotype (Figure S9E). This is consistent with its definition as a distinct disease type independent of ETS status.

Classification by evotype could have epidemiological implications. For instance, non-Caucasian racial groups display an increased incidence of many Alternative-evotype aberrations[29,30,31], and may therefore have a higher predisposition for this disease type. Conversely, cancers arising in younger patients have enrichment for ARBS-proximal breakpoints[18], and are reported to develop via a similar evolutionary progression to the Canonical-evotype[18,15]. It may also be possible to tailor treatment strategies to each evotype. In particular, cancers with aberrations found more commonly in the Alternative-evotype have been shown to be susceptible to ionising radiation[24], and have a better response to treatment with PARP inhibitors[32] and androgen ablation[25].

Our evolutionary model for prostate cancer disease types provides a conceptual framework that unifies the results of many previous studies and has significant implications for our understanding of progression, prognosis and treatment of this disease. As evolution through the sequential acquisition of synergistic genetic alterations is a process common to many tumours, the principles, analytical approach and conceptual framework outlined here are widely applicable and we anticipate them leading to insights into disease behaviour in other cancer types.

3.1. Limitations of Study

In this study we present evidence supporting the existence of at least two distinct evolutionary paths in prostate cancer, which underpins the concept of classifying these cancers into evotypes. However, the precise criteria that differentiate Canonical-evotype tumours from those of the Alternative-evotype remain to be rigorously defined. Our statistical classification may therefore have incorrectly assigned some tumours to an evolutionary path that does not reflect their true nature. Additionally, there is the possibility that a single prostate may contain tumour cell subpopulations following both trajectories. Although there was no evidence for this in the data sets we analysed, the most appropriate way to classify such cases remains undetermined. It is also likely that there are other evolutionary paths yet to be discovered, and so assigning these tumours to either of the two evotypes we describe here is incorrect. Another caveat is that our patient cohort predominantly consists of men of White-European ancestry treated in the UK, Australia and Canada, and therefore does not represent the global population. Therefore, while our findings are robust within the context of our study population, caution is warranted when extrapolating these results to other ethnic groups.

4. Acknowledgements

Funding. We acknowledge support from Cancer Research UK C5047/A29626, C5047/A22530, C309/A11566, C368/A6743, A368/A7990, C14303/A17197, the Dallaglio Foundation, Prostate Cancer UK and Prostate Cancer Research. We also acknowledge NIHR support to The Biomedical Research Centre at the Institute of Cancer Research and the Royal Marsden NHS Foundation Trust; Cancer Research UK funding to The Institute of Cancer Research and the Royal Marsden NHS Foundation Trust CRUK Centre; the National Cancer Research Institute (National Institute of Health Research (NIHR) Collaborative Study: "Prostate Cancer: Mechanisms of Progression and Treatment (PROMPT)" (grant G0500966/75466). We thank the National Institute for Health Research, Hutchison Whampoa Limited, University of Cambridge, the Human Research Tissue Bank (Addenbrooke's Hospital) which is supported by the NIHR Cambridge Biomedical Research Centre, and The Core Facilities at the Cancer Research UK Cambridge Institute. The Cambridge Human Research Tissue Bank and A.Y.W. are supported by the NIHR Cambridge Biomedical Research Centre. C.E.M was supported by a CRUK Major Centre Award through the CRUK Cambridge Centre Early Detection Programme and Urological Malignancies Programme. A.J.G. was funded by an SNF postdoctoral mobility fellowship (P2BSP3_178591). J.H.R.F. was supported by a Cancer Research UK Programme Grant to Simon Tavaré (C14303/A17197), as partially was A.G.L. A.G.L. acknowledges the support of the University of St Andrews. A.G.L. and J.H.R.F. also acknowledge the support of the Cambridge Cancer Research Fund. This work was supported by The Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2008), the UK Medical Research Council (CC2008) and the Wellcome Trust (CC2008) (M.T. and P.V.L.). For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission. M.T. was a postdoctoral fellow supported by the European Union's Horizon 2020 research and innovation program (Marie Skłodowska-Curie Grant agreement no. 747852-SIOMICS). P.V.L. is a CPRIT Scholar in Cancer Research, acknowledges CPRIT grant support (RR210006), and was a Winton Group Leader in recognition of the Winton Charitable Foundation's support towards the establishment of The Francis Crick Institute. G.S.B. was supported by the Academy of Finland; Cancer Society of Finland, and Sigrid Juselius Foundation. The Melbourne Prostate Cancer Research Group was supported by NHMRC project grants 1024081 (N.M.C. and C.M.H.) and 1047581 (C.M.H., N.M.C.), as well as a federal grant from the Australian Department of Health and Aging to the Epworth Cancer Centre, Epworth Hospital (A.J.C., N.M.C., C.M.H.). B.J.P. was supported by a Victorian Health and Medical Research Fellowship. The Canadian Prostate Cancer Genome Network was supported by Prostate Cancer Canada with funds from Movember and by the Ontario Institute for Cancer Research. P.C.B. was supported by the NIH/NCI under award number P30CA016042. V.B. was funded by a Fellowship from the Canadian Institutes for Health Research and was supported by a Michael Smith Foreign Study Supplement from the Canadian Institutes for Health Research. We also acknowledge support from the Bob Champion Cancer Research Trust, the Masonic Charitable Foundation, the King Family and the Stephen Hargrave Foundation. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The authors would like to thank those men with prostate cancer and the subjects who have donated their time and samples for this study.

Author Contributions. D.J.W., C.S.C. and D.C.W. developed the concepts in the study. D.C.W. initiated and supervised the study. Analysis performed by D.J.W. and A.S.. Manuscript was written by D.J.W., C.S.C. and D.C.W. with contributions and revisions from all authors. STAR Methods written by D.J.W., A.S., A.Z., and D.S.B. with contributions from all authors. UK sample collection and clinical data generation overseen by V.G., A.W., F.C.H., G.S.B, C.S.F., D.E.N, Y.-J. L., D.S.B., and R.A.E.. UK WGS data processing performed by G.G., Y.X., M.G., A.B., and D.S.B. Data from Australia generated and provided by B.J.P., C.-H.J., A.J.C., N.E.C. and C.M.H. Data from Canada provided by V.B., M.F., R.G.B and P.C.B.. Bioinformatics measurements created by D.J.W., A.S., R.T., C.E.M., A.J.G., V.B., E.A., M.T., S.D. and J.H.R.F., with oversight from P.V.L., Z.K.-J., P.C.B., A.G.L. and D.C.W.. Data generation, standardisation and storage overseen by D.S.B.. R.A.E. and C.S.C. co-directed the project. The full CRUK ICGC Prostate Group created and maintains overall project direction; principal investigators are G.S.B, C.E.M., A.G.L., D.S.B., R.A.E., C.S.C., D.C.W.

Competing interests. An international patent related to this work has been published under International Publication Number WO 2023/047140 A1. R.E. has the following conflicts of interest to declare: Honoraria from GU-ASCO, Janssen, University of Chicago, Dana Farber Cancer Institute USA as a speaker. Educational honorarium from Bayer and Ipsen, member of external expert committee to Astra Zeneca UK. She undertakes private practice as a sole trader at The Royal Marsden NHS Foundation Trust and 90 Sloane Street SW1X 9PQ and 280 Kings Road SW3 4NX, London, UK. All other authors declare no competing interests.

5. Figure legends

Figure 1. Co-occurrence of genetic alterations distinguishes three metaclusters. After performing feature extraction, we calculated a discrimination score quantifying the relevance of each feature in predicting relapse (green heatmap). Fourteen features (red) were used as inputs for k-medoid clustering with 11 clusters. The medoids of each cluster were used as inputs to hierarchical clustering using all features, which revealed three main metaclusters, MC-A, MC-B1 and MC-B2, with different profiles as indicated by the dendrogram. The main heatmap shows the medoid feature values for the patients in each cluster, ordered by the hierarchical clustering (scale to right). The number of samples in each cluster is given below the corresponding cluster medoid. Metacluster colours are denoted by text above dendrogram.

Figure 2. Classification by proximity of DNA breakpoints to AR binding sites reveals common genetic alterations. (A) The proportion of DNA breakpoints within 20 kilobases (kb) of an AR binding site for each patient, normalised by the number of proximal breakpoints expected by chance (vertical axis). Tumour samples are ordered according to this normalised proportion (horizontal axis). Classes were determined based on whether the tumour displayed more (Enriched) or fewer (Depleted) proximal breakpoints than expected, or there was no statistical significance (Indeterminate). (B) Heatmaps of genomic features for each patient, ordered as above. Statistically significant relationships for the three classes are shown in the 'Relationship' column, where E, D and I indicate the Enriched, Depleted and Indeterminate classes respectively. Braces indicate no relationship between the enclosed classes, but they both display significant differences to the remaining class. Relationships are ordered so the leftmost class(es) are those showing significantly greater proportion of the corresponding genetic alteration. For Bernoulli variables, significance was determined with Chi-squared test followed by a Fisher exact test for each pairwise relationship, for continuous variables a Kruskal-Wallace test with Tukey's HSD was used (FDR adjusted $p < 0.05$ for all tests).

Figure 3. Samples can be differentiated by order of genetic alterations. Phylogenetic trees from individual tumours were used to estimate two ordering profiles using a Plackett-Luce (P-L) mixture model. Tumours were assigned to Ordering-I (top) or Ordering-II (bottom). Horizontal box and whisker plots (5th/25th/75th/95th percentiles) represent the spread of bootstrap estimates of the negative Plackett-Luce coefficient (α_i) for the ith genetic alteration (x-axis). Here, the lower the value of α_i, the earlier the genetic alteration is likely to occur. The y-axis shows the proportion of samples in the mixture component in which the genetic alteration was observed. Colours of the box and whiskers denote the chromosome on which the aberration occurred. Genetic alterations were annotated if they were identified as an ETS fusion, occurred with a proportion above 0.25 or were identified in the earliest 5 events; these have chromosomal regions given with notable driver genes in the region given in brackets where applicable. Other genetic alterations were not annotated and are displayed with reduced transparency.

Figure 4. Integrating results reveal multiple evolutionary trajectories converging to two disease types with different prognosis. (A) A comparison of how tumours were classified in each of the three previous methods. Each side of the triangle corresponds to a classification method, wherein each bar in the triangle denotes a group identified by that method. Values at the intersections of each bar show the number of tumours which were consistent to both classes. Values outside the main triangle denotes the total number of tumours in that class. Colours are those used in previous figures. (B) A schematic of the evotype model for prostate cancer evolution. (C) The prevalence of each genetic aberration in each evotype, as determined using the majority consensus of the three classifiers. Aberrations with significant differences between evotypes are coloured by the evotype displaying the highest proportion (FDR adjusted $p < 0.05$, Fisher Exact Test). (D) A surface plot showing the probability density of a tumour being assigned to the Canonical-evotype relative to the number of aberrations. Common modes of evolutionary progression follow regions of high density as the number of aberrations increase. Exemplars of such routes are indicated by black dashed lines. These are labelled according to their likely evotype, a behavioural descriptor, and notable driver genes affected by aberrations that are prevalent in the areas along the path to convergence (Figures S7 and S8).

Figure 5. Frequency of AR-induced DNA loops associated with DSBs is associated with *CHD1* loss and evotype status. A) A simplified schematic of AR binding to AR-binding sites (ARBS), where CHD1 protein is part of a complex that induces DNA loop formation and subsequent DSBs, denoted by the red X. B) A notched box and whisker plot shows that adjacent proximal ARBS pairs that are required for DNA loops to form were observed less frequently in the vicinity of breakpoints in *CHD1*-deficient tumours than *CHD1* wildtype tumours. C) DSB-associated ARBS pairs also occurred less frequently in tumours of the Alternative-evotype than the Canonical-evotype, even when *CHD1* is unaffected. D) P-values were determined through a one-sided Mann-Whitney U-test.

Figure 6. Utility of evotype model in survival analysis. Kaplan-Meier plots for: (A) the evotypes, (B) 20 tumours with greatest tumour mutational burden (High TMB) against the remainder (Low TMB), (C) ISUP Gleason grade, (D) 10 tumours with highest TMB for each evotype (High-TMB Alternative and High-TMB Canonical) against the remainder (Low-TMB Combined), (E) The ISUP Gleason grade ≥ 3 tumours in the High TMB evotype classes (Evo-TMB-Gleason High) and the remainder (Evo-TMB-Gleason Low), (F) Alternative evotype tumours in MC-A (MC-A/Alternative), Canonical-evotype tumours in MC-B2 (MC-B2/Canonical) and the remainder (MC-B1/Combined), and (G) ISUP Gleason grade ≥ 3 tumours of either MC-A/Alternative or MC-B2/Canonical (MC-A/B2-Gleason High Combined) against the remainder MC-A/B1/B2-Gleason Low Combined. For each comparison we provide the hazard ratio (HR) and p-value calculated with Cox proportional hazard test, p-value adjusted for Gleason grade, TMB and age-at-diagnosis if they are not used to create the sets used in the comparison (p_{adj}), and Harrell's c-index. In D and F, these values are given for the denoted class in comparison to the remainder

only. End point is time to biochemical recurrence.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact. Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, David C. Wedge, Ph.D. (david.wedge@manchester.ac.uk)

Materials availability. There are no tangible materials produced by this study that are available for distribution.

Data and code availability. Sequencing data generated for this study have been deposited in the European Genome-phenome Archive with accession code EGAS00001000262. Processed data and code used in this manuscript is available at `https://github.com/woodcockgrp/evotypes_p1/` and via `https://doi.org/10.5281/zenodo.10214795`.

Experimental model and study participant details

Cancer samples from radical prostatectomy, and matched blood controls, were collected from 205 patients treated at the Royal Marsden NHS Foundation Trust, London, at the Addenbrooke's Hospital, Cambridge, at Oxford University Hospitals NHS Trust, and at Changhai Hospital, Shanghai, China, as described previously[33,34]. Ethical approval was obtained from the respective local ethics committees and from The Trent Multicentre Research Ethics Committee. All patients were consented to ICGC standards. 159 of the samples passed stringent quality control for copy number profiles and structural variants, and were used in this study.

METHOD DETAILS

DNA preparation and DNA sequencing

DNA from frozen tumour tissue and whole blood samples (matched controls) was extracted and quantified using a ds-DNA assay (UK-Quant-iT™ PicoGreen® dsDNA Assay Kit for DNA) following the manufacturer's instructions with a Fluorescence Microplate Reader (Biotek SynergyHT, Biotek). Acceptable DNA had a concentration of at least 50ng/μl in TE (10mM Tris/1mM EDTA), and displayed an optical density 260/280 (OD_{260}/OD_{280}) ratio between 1.8-2.0. Whole Genome Sequencing (WGS) was performed at Illumina, Inc. (Illumina Sequencing Facility, San Diego, CA USA) or the BGI (Beijing Genome Institute, Hong Kong), as described previously[33,34], to a target depth of 50X for the cancer samples and 30X for matched controls[33]. The Burrows-Wheeler Aligner[35] (BWA) was used to align the sequencing data to the GRCh37 reference human genome.

Generation of summary measurements

We generated 123 summary measurements from the WGS data using a number previously published algorithms, so we briefly outline those below. These are grouped into measurements that were generated with similar or related algorithms; default parameters were used unless otherwise stated. The processed data is given alongside the code at `https://github.com/woodcockgrp/evotypes_p1/`.

Numbers of SNVs, indels and structural variants - 10 fields. SNVs, insertions and deletions were detected using the Cancer Genome Project Wellcome Trust Sanger Institute pipeline as described previously[33]. In brief, SNVs were detected using CaVEMan with a cut-off 'somatic' probability of 0.95. Insertions and deletions were called using a modified version of Pindel[36]. Variant allele frequencies of all indels were corrected by local realignment of unmapped reads against the mutant sequence. Structural variants were detected using Brass[33]. Total numbers of SNVs, indels and rearrangements per sample were calculated (1 field each), as were types of indel (3 fields: insertion, deletion and complex) and structural variants (4 fields: large insertions or deletions, tandem duplications and translocations).

Percentage genome altered - 3 fields. This was calculated as the percent total of the genome that is affected by CNAs[37]. We also recorded the percentage affected by clonal and subclonal CNAs (i.e. CNAs with CCF=1 and CCF<1 respectively).

Ploidy - 1 field. We adopt the same approach as detailed previously[11], where whole genome duplicated samples were those which had an average ploidy, as identified with the Battenberg algorithm, greater than 3. These samples were designated as tetraploid and assigned a value of 1 in our data set, otherwise the sample was diploid (assigned 0).

Kataegis - 1 field. Kataegis was identified using SeqKat `https://github.com/cran/SeqKat`. The datum was set to 1 if kataegis was identified and 0 if not.

ETS status - 1 field. A positive ETS status was assigned if a DNA breakpoint involving *ERG*, *ETV1*, *ETV3*, *ETV4*, *ETV5*, *ETV6*, *ELK4*, or *FLI1* and partner DNA sequences was detected and the fusion was in-frame. The datum was set to 1 if there was ETS fusion detected or 0 if not.

Gene fusions - 2 fields. We reported the number of in-frame gene fusions in the sample (counts) and if there was a gene fusion affecting the *TMPRSS2*/*ERG* genes (1 or 0)

Breakpoints - 14 fields. Breakpoints were identified with Chainfinder[38] version 1.01. Total number of breakpoints, total number of chained breakpoints (i.e. where the breakpoints are interdependent), number of chains, the number of breakpoints in the longest chain, the number of breakpoints involved in the chained events, and the maximum number of chromosomes involved in a chain were recorded as integer counts (6 fields). We also calculated the proportion of all breakpoints that were in chained events (1 field - $[0, 1]$) and the average, median and maximum number of chromosomes involved in a chain (3 fields - $[0, \infty]$). Information about the type of breakpoint was also recorded, including the number of deletion bridges, intra-chromosomal and inter-chromosomal events (3 fields - counts) and the inter-chromosomal to intra-chromosomal ratio (1 field - $[0, \infty]$, set to zero if there were no intra-chromosomal breakpoints).

Mutated driver genes - 26 fields. A set of driver genes were identified from our previous publication[11]. Using the CaVEMan output, we determined any non-synonymous mutations in the exonic regions of these genes as a mutated driver gene; the corresponding field was assigned a value 0 if no such mutations were identified and 1 if there were.

Copy number alterations - 60 fields. We followed our previous approach[11] to identify consistently aberrant regions. A permutation test was developed where CNAs detected from each sample were placed randomly across the genome and then the total number of times a region was hit by each type of CNA in this random assignment was compared to the number of times a region was hit in the actual data. This process was repeated 100,000 times and recurrent (or enriched) regions were defined as having a false discovery rate (FDR) of less than 0.05. This was performed separately for gains, loss of heterozygosity (LOH) and homozygous deletions (HD). We identified small regions initially and these were amalgamated into larger regions defined as the regions between chromosomal positions when the difference between the number of CNAs identified in the data and expected frequency (if this process were uniformly random) dropped to zero. For each sample, if a breakpoint corresponding to a gain, LOH or HD occurred in each region, then the respective datum was set to 1, and 0 otherwise.

Telomere lengths - 1 field. Telomere lengths were estimated as described in our previous publication[39]. A mean correction was applied to batches to compensate for the effects of a change in chemistry during the project, therefore the value is continuous in the range $[0, \infty]$.

Chromothripsis - 4 fields. The identified copy number breakpoints were segmented in inter-breakpoint distance along the genome using piecewise constant fitting (`pcf` from the R package `copynumber` v1.22.0). Regions with a density higher than 1 breakpoint per 3Mb were flagged as high-density regions. A chromothripsis region was then defined as a high-density region with a number of copy number breakpoints $N > 15$; a non-random segment size distribution (Kolmogorov-Smirnov test against the exponential distribution,

$P < 0.05$); at most three allele-specific copy number states covering more than $\min(1, -0.006N + 1.1)$ fraction of the region; and the proportion of each type of structural variant is random with equal probability $P_{\mathrm{TD}} = P_{\mathrm{Del}} = P_{\mathrm{H2Hi}} = P_{\mathrm{T2Ti}} = 0.25$ (multinomial test $P > 0.01$), where TD=tandem duplication, Del=deletion, H2Hi=head-to-head inversion and T2Ti=tail-to-tail inversion. We recorded the presence or absence of chromothripsis (1 or 0 respectively), the proportion of all breakpoints in chromothripsis events ($[0, 1]$), the number of chromothripsis events in each sample (counts) and the size of the largest chromothripsis region (counts).

QUANTIFICATION AND STATISTICAL ANALYSIS

In this section we aim to provide a largely non-technical overview of each of our methods we used to perform the analysis in the study, followed by more technical description for those who wish to fully understand and reproduce our methodology.

Statistics

Prior to the study we predetermined we would use Fisher's Exact Test for 2x2 contingency tables and Chi-squared test for contingency tables of greater dimensionality and this is applied throughout. Associations between genetic alterations and ARBS clusters was identified using one-tailed Fisher Exact Test with p <0.05, corrected for multiple testing using the False Discovery Rate. Relationships were determined dependent on the variable type: for Bernoulli variables, significance was determined with Chi-squared test followed by a one-tailed Fisher exact test for each pairwise relationship; one-tailed tests were used as a two-tailed test would not have revealed the direction of the relationship. For continuous variables a Kruskal-Wallace test with Tukey's HSD was used (adjusted p <0.05 for all tests). Significance of Depleted groups across countries clustering together was determined using the Approximately Unbiased Multiscale Bootstrap procedure. Associations between evotypes and individual genetic alterations was conducted with a two-tailed Fisher Exact Test, corrected for multiple testing using the False Discovery Rate. The associations with ARBS pairs were established with a one-sided Mann-Whitney U-test with p <0.05. Statistics associated with the Kaplan-Meier plot were calculated using log-rank methods, and significance level was set at 0.05. Cramer's V statistic was used to determine the strength of the associations in between the cluster assignments. As we only claim an association between patients assigned to MC-A (Metaclusters), the Depleted group (ARBS) and Ordering II, we combined metaclusters MC-B1 and MC-B2 into one class and the Enriched and Depleted ARBS groups into one class for this comparison.

Unsupervised feature extraction

The summary measurements detailed above form the data set for further analysis. However, it contains a number of different data types (binary, proportions, continuous, integer counts), it is high dimensional relative to the number of patients, and it undoubtedly contains highly correlated, cooccurring or equivalent events that may confound the analysis. To address this we performed a feature extraction preprocessing step prior to the analysis. As our downstream analysis will be investigating genomic patterns that are indicative of evolutionary behaviour, it is critical that the results of these analyses can be easily interpreted. This necessitates methodology where the links between input variables that correspond to the features are identifiable. We therefore opted for a *latent feature* approach as the basis of our feature extraction as these can provide an interpretable representation of the relationships between the inputs[40]. Latent feature (or latent variable) analysis provides a way of reformulating the data into a reduced set of *features* that encapsulate the underlying relationships between the original inputs. The data can be recast in terms of these latent features, which is known as the *latent feature representation*, and the downstream analysis performed directly on this.

There have been many latent feature models proposed, each with associated inference methods for the features (a process called *feature learning*). These included methods such as non-negative matrix factorisation[41], Bayesian nonparametric methods[42] and neural networks[43]. However, none of these were able to fulfil all of our requirements above. We therefore created a bespoke method for feature extraction on this data set.

Neural networks for feature extraction

We utilised a Restricted Boltzmann Machine[44] (RBM) neural network as the basis of our feature learning method. We chose to use an RBM as it is extensible to multiple data types[45,46] and can provide interpretable hidden units, with appropriate modifications[47]. An RBM is functionally similar to another type of neural network architecture called an autoencoder[43]. Autoencoders are a class of network types that compress (*encode*) the data into a transformed representation (the *code*), and then decompress (*decode*) in an attempt to reconstruct the original data. A measure of the error between the reconstruction and the original data is used to update the parameters through backpropagation. Typically the code layer contains fewer units than the input/output layers and this bottleneck means that the learning process attempts to compress the information in the data set into a more a compact representation in the code.

In contrast, the basic RBM unit consists of only two layers, known as the *visible* and the *hidden* layers. The RBM is formulated as a probabilistic network, meaning each unit represents a random variable rather than a fixed value. As such, the hidden layer performs a similar function to the code layer in the autoencoder, albeit with a probabilistic representation. It has been shown that the RBM is equivalent to the graphical model of factor analysis[48] and so each hidden unit can be interpreted as a latent feature. Another distinction from the autoencoder formulation is that there is only one weight matrix, which used to update both the visible and hidden layers. This means that the information on the transformation from visible units (input representation) to the hidden units (feature representation) is encapsulated in this matrix. Hence we also refer to it as the input-feature map.

The Restricted Boltzmann Machine

The standard RBM formulation[44] consists of Bernoulli random variables for all visible $\mathbf{v} = \{v_i\}$ and hidden units $\mathbf{h} = \{h_i\}$, where $v_i, h_j \in \{0, 1\}$, with respective biases $\mathbf{a} = \{a_i\}, \mathbf{b} = \{b_j\}; a_i, b_j \in (-\infty, \infty)$, and a matrix of weights, $W; w_{ij} \in (-\infty, \infty)$. Training of an RBM is based on minimising the *free-energy* of the visible units, as a low free-energy corresponds to a state where the data is explained well through the model parameterisation. Energy-based probability distributions take the form

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}, \tag{1}$$

where $E(\mathbf{v}, \mathbf{h})$ is the energy function and Z is a normalising factor. This is the probability of observing the joint $\mathbf{v}, \mathbf{h}$ pair. The energy function in an RBM is given as

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{a}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{v}^T W\mathbf{h}. \tag{2}$$

In this formulation,

$$Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}, \tag{3}$$

which is difficult to calculate due to the number of possible combinations of $\mathbf{v}$ and $\mathbf{h}$.

As we want training to be conducted with respect to the energy at the visible units, we need to marginalise over $\mathbf{h}$ in Equation 1 to calculate the likelihood of observing the visible unit corresponding to a single data sample $\mathbf{d}_k$ from data set $D = \{\mathbf{d}_k, k = 1, 2, \dots, K\}$.

$$\mathcal{L}(\theta|\mathbf{v} = \mathbf{d}_k) = \frac{1}{Z} \sum_{\mathbf{h}} e^{-E(\mathbf{d}_k,\mathbf{h})}, \tag{4}$$

where $\theta \in \{\{a_i\}, \{b_j\}, \{w_{ij}\}\}$ is the full parameter set. To simplify notation, we write $\mathcal{L}(\theta|\mathbf{v} = \mathbf{d}_k)$ as $\mathcal{L}(\mathbf{d}_k)$ with no loss of generality. To perform training through gradient descent, we need to calculate the gradient of the negative log-likelihood for each parameter we wish to update, $\partial(-\log \mathcal{L}(\mathbf{d}_k))/\partial\theta$. The partial derivative of the logarithm of Equation 4 takes the form

$$\begin{aligned}
\frac{\partial}{\partial\theta}(-\log \mathcal{L}(\mathbf{d}_k)) &= \frac{\partial}{\partial\theta}\left(\log \sum_{\mathbf{h}} e^{-E(\mathbf{d}_k,\mathbf{h})}\right) - \frac{\partial}{\partial\theta}\left(\log \sum_{\mathbf{v}} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}\right) && (5)\\
&= \sum_{\mathbf{h}} P(\mathbf{h}|\mathbf{v} = \mathbf{d}_k) \frac{\partial E(\mathbf{d}_k, \mathbf{h})}{\partial\theta} - \sum_{\mathbf{v}} \sum_{\mathbf{h}} P(\mathbf{v}, \mathbf{h}) \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta} && (6)
\end{aligned}$$

We then calculate the expected values using the entire training set

$$\mathbb{E}_D\left[\frac{\partial}{\partial\theta}(-\log \mathcal{L}(\mathbf{d}_k))\right] = \mathbb{E}_{P(\mathbf{h}|D)}\left[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}\right] - \mathbb{E}_{P(\mathbf{v},\mathbf{h})}\left[\frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial\theta}\right], \tag{7}$$

which can be used to update the model parameters via gradient descent. The $\mathbb{E}_{P(\mathbf{h}|D)}$ term corresponds to the expected energy state invoked from observing the data samples, and the $\mathbb{E}_{P(\mathbf{v},\mathbf{h})}$ is the expected energy state of the model configurations, both contingent on the current model parameters. As such, they are often called $\mathbb{E}_{data}$ and $\mathbb{E}_{model}$ respectively. Calculating the partial derivatives with respect to the parameters gives

$$\begin{aligned}
\frac{\partial}{\partial w_{ij}}(-\log \mathcal{L}(\mathbf{d}_k)) &= \mathbb{E}[v_i h_j | \mathbf{v} = \mathbf{d}_k] - \mathbb{E}[v_i h_j], && (8)\\
\frac{\partial}{\partial a_i}(-\log \mathcal{L}(\mathbf{d}_k)) &= \mathbb{E}[v_i | \mathbf{v} = \mathbf{d}_k] - \mathbb{E}[v_i], && (9)\\
\frac{\partial}{\partial b_j}(-\log \mathcal{L}(\mathbf{d}_k)) &= \mathbb{E}[h_j | \mathbf{v} = \mathbf{d}_k] - \mathbb{E}[h_j], && (10)
\end{aligned}$$

which are used to construct the update equations

$$\begin{aligned}
W^{new} &\leftarrow W^{old} + \nu\big(\mathbb{E}_{data}[\mathbf{v}^T\mathbf{h}] - \mathbb{E}_{model}[\mathbf{v}^T\mathbf{h}]\big), && (11)\\
\mathbf{a}^{new} &\leftarrow \mathbf{a}^{old} + \eta\big(\mathbb{E}_{data}[\mathbf{v}] - \mathbb{E}_{model}[\mathbf{v}]\big), && (12)\\
\mathbf{b}^{new} &\leftarrow \mathbf{b}^{old} + \eta\big(\mathbb{E}_{data}[\mathbf{h}] - \mathbb{E}_{model}[\mathbf{h}]\big), && (13)
\end{aligned}$$

for learning rates ν and η. The $\mathbb{E}_{data}$ values can be estimated easily by taking the arithmetic mean.

The $\mathbb{E}_{model}$ terms are generally difficult to calculate as they involve summation over all possible configurations of $\mathbf{v}$ and $\mathbf{h}$. An alternative is to perform Gibbs sampling using the conditional probabilities as these are far easier to calculate due to the conditional independence between units in the same layer. We can estimate the conditional probability of values of the hidden layer from the visible layer and vice versa thus

$$\begin{aligned}
P(\mathbf{h}|\mathbf{v}) &= \prod_j P(h_j|\mathbf{v}), && (14)\\
P(\mathbf{v}|\mathbf{h}) &= \prod_i P(v_i|\mathbf{h}). && (15)
\end{aligned}$$

The form of $P(h_j|\mathbf{v})$ and $P(v_i|\mathbf{h})$ depends on the *activation function.* This function that inputs the products of the units in one layer and their corresponding weights, and outputs a probability that a unit is active. In this study, we use a *logistic sigmoid* (or simply "sigmoid") function, which is given by

$$\sigma(x) = \frac{1}{1+e^{-x}}, \qquad (16)$$

where x is dependent on the layer we are sampling, and so the individual hidden and visible probabilities can be written as

$$\begin{aligned}
P(h_j|\mathbf{v}) &= \sigma(b_j + \sum_i v_i w_{ij}), && (17)\\
P(v_i|\mathbf{h}) &= \sigma(a_i + \sum_j h_j w_{ij}). && (18)
\end{aligned}$$

A sample is drawn by setting the corresponding unit to 1 with probability given by the value for $P(h_j|\mathbf{v})$ or $P(v_i|\mathbf{h})$ as appropriate. These can then be used to calculate estimates for $P(\mathbf{v})$ and $P(\mathbf{h})$ by marginalisation over the conditional variable. In practice a full Gibbs sample every update iteration would be prohibitively slow and so we used an approximation called *contrastive divergence*[44], in which the Gibbs sampler is initialised using the input data and a limited number of Gibbs steps are performed. In our implementation we use one contrastive divergence step (i.e. $CD(1)$), and so the data (or mini-batches of the data) is presented as a matrix and used to sample the hidden unit values, which are then used to update the values of the visible units. These values are used to update the network parameters using stochastic gradient descent (SGD)[49].

During training, the results of these updates are stored in three matrices that correspond to the weights as well as the network representation of the tumour data at the visible and hidden layers. These matrices correspond to the network reconstruction of the data (visible layer, $\mathbf{V}$) the latent feature representation of the data (hidden layer, $\mathbf{H}$), and the input-feature mapping (weights, $\mathbf{W}$). When the network is trained, these can be extracted and utilised in the analysis.

Modifications to the base RBM

We made a number of simple modifications to the base RBM described above to ensure the feature representation was interpretable, generalisable, stable and reproducible. These modifications are described below.

Data Integration. Our data consisted of multiple different modalities; unlike conventional multiomics approaches which have a large number of a data points from a small number of sources, we have a small number of data points from a large number of sources. As such, data integration needed to be carefully considered. The RBM can be modified to incorporate inputs of multiple modalities, sometimes through modification of the energy function[50,51]. However, we decided to avoid this complication and standardise all our inputs by ranking all integer and continuous variables prior to rescaling to $[0, 1]$. Specifically, our transformations were

- Binary – set as $\{0, 1\}$,
- Integer – rank and scale to $[0, 1]$,
- Continuous – rank and scale to $[0, 1]$.

For the integer and continuous cases we used ranking as this decouples the value from the distribution of the inputs and after scaling to $[0, 1]$, the new value can be interpreted as the probability that the corresponding visible unit is active. As such, all inputs are treated equally in the machinations of the RBM. These transformations do not affect the hidden units, which remain a Bernoulli random variable, $h_i \in \{0, 1\}$.

Non-negative weights. Neural networks are considered as black-box approaches as the transformations they perform are highly complex. To improve interpretability of the network machinations we imposed a non-negativity constraint to the weight updates, specifically by penalising negative values. We use an approach in which a quadratic barrier function is subtracted from the likelihood for each negative weight[47]. Mathematically, this is written as

$$\mathcal{L}(\mathbf{d}_k)_{nonneg} = \mathcal{L}(\mathbf{d}_k) - \frac{\alpha}{2} \sum_i \sum_j f(w_{ij}), \tag{19}$$

where α denotes the strength of the penalty, and

$$f(x) = \begin{cases} x^2, & \text{if } x < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

This leads to the update rule

$$W^{new} \leftarrow W^{old} + \nu\big(\mathbb{E}_{data}[\mathbf{v}^T\mathbf{h}] - \mathbb{E}_{model}[\mathbf{v}^T\mathbf{h}] - \alpha W^{\{\text{-}\}}\big), \tag{21}$$

where $W^{\{\text{-}\}}$ is a matrix containing the negative entries of W, with zeros elsewhere. This formulation is equivalent to a L_2-norm penalty on the negative weights, and so penalises more strongly negative weights to a greater degree. When used in the training scheme, this coerces network weights to non-negative solutions, simplifying the interpretation of the input-feature map. This can be considered to be a non-linear extension of non-negative matrix factorisation[41], and similarly can be used to represent the underlying structure of the data by its *parts*, which is synonymous with latent features here.

As weights can no longer trade off against each other with counteracting weights of opposing signs, this means that the lowest free-energy state corresponds to a state with minimal redundancy and so during training the hidden units compete to convey information about a single input[52]. This means that the input will only be represented in small number of latent variables, so when the initial number of hidden units is of similar order to the number of data inputs, this results in some of the biases or weights converging to a negligible value, and the corresponding hidden layer activations converge to an arbitrary fixed value. The latter are then called *dead units.* This is of fundamental importance to our method as it can be used as an estimate of the intrinsic dimensionality of the data.

Hidden unit pruning. During training, we prune the *dead units* to improve the speed of the algorithm. However determining dead units is not straightforward in a probabilistic network such as the RBM as the values in the network at each state will vary stochastically. To circumvent this, we apply an $L_{1/2}$-norm penalty on the hidden unit activations, which penalise a non-zero activation value[53]. This coerces the values for all patient samples to be zero, rather than some arbitrary value, and these can then be easily identified and removed with a thresholding approach. This penalty function is calculated over all training data samples, so for consistency with Equation 4 we can formulate the likelihood for each sample as

$$\mathcal{L}(\mathbf{d}_k)_{activ} = \mathcal{L}(\mathbf{d}_k) - \frac{\beta}{K} \sum_k \|f(y_k)\|_{1/2}\,, \tag{22}$$

where $f(y_k) = P(\mathbf{h}|\mathbf{v}_k)$ and β is a parameter describing the strength of this penalty. We calculate the gradient of the additional likelihood term with respect to each of the hidden unit biases, which is given as

$$\begin{aligned} \Delta b_j^{(L_{1/2})} &= \frac{1}{K} \frac{\partial \sum_k \|P(\mathbf{h}|\mathbf{v}_k)\|_{1/2}}{\partial b_j}, & (23) \\ &= \frac{1}{2} \sum_k \frac{\exp(-b_j - \sum_i v_{ik} w_{ij})}{|1 + \exp(-b_j - \sum_i v_{ik} w_{ij})|^{3/2}}. & (24) \end{aligned}$$

We can then write the vector of gradients for all hidden unit biases as $\Delta \mathbf{b}^{(L_{1/2})}$. The corresponding update rule can therefore be written as

$$\mathbf{b}^{new} \leftarrow \mathbf{b}^{old} + \eta\big(\mathbb{E}_{data}[\mathbf{h}] - \mathbb{E}_{model}[\mathbf{h}]) - \beta \Delta \mathbf{b}^{(L_{1/2})}. \tag{25}$$

In our training algorithm, we prune dead units every 50 iterations after the first 1000 iterations.

Sparsity. Sparsity is a desirable property for latent space representations, as it means that the information is conveyed in a concise form. The penalty measure defined in Equation 22 introduces sparsity as it penalises hidden units which are highly active thus coercing the network toward a sparse configuration[53]. Further sparsity measures were not used in training as the weight matrix, which defines the input to feature mapping, will be filtered at a later stage.

Overfitting. A concern with any neural network formulation is the tendency to overfit the data, which in this application would lead to a feature set that was not representative of the true underlying structure, and therefore not generalisable. To mitigate this, we employed a number of countermeasures, namely

1. DropConnect,
2. Max-norm regularisation,
3. Bootstrap aggregating,
4. Early Stopping.

With DropConnect[54], a predetermined proportion of weights in the network are randomly set to zero with uniform probability at each training iteration. This helps prevent overfitting by temporarily disrupting correlations between features, so they are more likely to learn features that are independent of the state of other features.

When using max-norm regularisation[55], we set an absolute value on the norm of each weight vector that form the input to a single hidden unit. If a vector becomes too large then we rescale the vector so that it obeys the constraint.

It is possible for non-negative weights to continue increasing throughout training as the binary nature of some inputs means that when present they were already in the maximal output of the sigmoid activation function so the precise value is irrelevant. Max-norm regularisation prevents this occurrence and facilitates comparison between weight matrices of different runs.

For bootstrap aggregating[56] (bagging), multiple networks with the same initial architecture were trained on subsets of the data and the outputs amalgamated. In our feature learning representation, we extracted the weight matrix from each of the networks and merged them according to the cosine distance between features.

Finally, when implementing early stopping[57] we need to compare the performance of the network on the training set to the performance on an unseen *validation set*. If the network performs similarly on the training and validation sets then it is a good indicator that it will return generalisable outputs. Beginning with the subsets extracted for ensemble learning, we use data omitted when the subset was sampled as the validation set, which is propagated through the network. As the RBM is formulated as an energy-based model, early stopping is predicated by comparing the *free energy* in the training set to that in the validation set[58]. In general overfitting-mitigation strategies, the free energy (or reconstruction error in error-based networks) is monitored and if the free energy in the training set decreases while the free energy in the validation set increases, that indicates overfitting is occurring and training is stopped. We adopted a more stringent approach in which the samples in the training set are randomly assigned to subsets of equal size to the validation set and so the free energy values are directly comparable. During training, if the free energy of the validation set increases above the largest free energy of the training subsets for an extended period (10 iterations) then training is stopped and the entire run is discarded and training repeated. This means that the network is able to model unseen data (the validation set) as well as it does the training set when accounting for variation in energy values resulting from sampling the validation set. If overfitting is suspected, the entire run is discarded and another training run performed; as our main objective is to derive the input-feature mapping via the weight matrix, this avoids the situation in which we retain a weight matrix that has not had time to converge to a solution consistent with those runs that completed without interruption.

Convergence to global solution. As we are training multiple networks and amalgamating the results, it is important that each network converges to the global solution or the results will be incongruous. Furthermore, as the RBM is trained by stochastic gradient descent, it is possible that the algorithm may get stuck in a local optima. To minimise the chance of this occurrence, we used the *cyclical learning rate* scheme[59], in which learning rates for each of the variables oscillates between zero and a maximal value throughout training. The maximal value is subject to decay so that the maximal training rate will diminish throughout training to zero. This approach has been shown to help convergence to the global solution and has the advantage that the learning rate parameters do not

```
input: set of weight matrices from each network run
concatenate weight matrices into matrix $W$;
set low magnitude weights to zero;
set similarity threshold $\tau = 0.5$;
Initialise matrix $M$ with number of rows and columns equal to number
  of inputs;
Initialise empty feature matrix $F$;
for i=1 to number of inputs do
    set $i^{th}$ row of $M$ equal to mean of all rows of $W$ where the $i^{th}$
      weight $> 0$
end
calculate pairwise cosine similarity matrix $S$ from $M$;
while number of rows in S¿0 do
    read in the first row of $S$ as the current similarity vector $\mathbf{s}$;
    identify all $j$ where $\mathbf{s}_j > \tau$;
    add the mean of all $M[j,:]$ as a row to $F$;
    remove $j^{th}$ rows from $S$ and $M$;
end
rescale all rows in feature matrix $F$ by max-norm;
```

Algorithm 1: Pseudocode for amalgamating weight matrices

need to be tuned[59].

Amalgamation of feature matrices

Each individual network run provides a similar, but not identical, weight matrix. As such, weight matrices from each network run were amalgamated and filtered to form the final input-feature map. Numbers of features, the inputs they represent, their magnitude and order would not necessarily occur the same in each network and so we constructed an algorithm based on the cosine similarity, is which depicted in Figure S10, and outlined in Algorithm 1. Note that *Low magnitude weights* were those less than 50% of the maximum weight value for each hidden unit.

Synthetic data

To investigate whether our RBM network can identify true associations in data of multiple types, we trained the network on a synthetic data set with known associations and data generation methods. The values in the synthetic data set were generated from function that encapsulated simple relationships when applied to binary latent variables; we also utilised various statistical distributions on top of these relationships to model types like proportions and counts that we might find in the real data set. The synthetic data is constructed in seven ‘blocks’ to aid interpretation. In the first six blocks, 5 latent variables are mapped to 10 observed variables in exactly the same way. The difference between the blocks is the statistical distributions used to generate the values in the data. The final block consists of latent variables mapped to distributions from all the previous types. In total there were 72 ‘observed variables’ generated from 34 ‘latent variables’. To generate the data for a single synthetic sample, the 34 binary latent variables were sampled uniformly with probability of the latent variable being 1 set at 0.5. This was done for 200 synthetic samples to create the synthetic data set.

Latent variable to observed variable mappings

We denote the ith simulated observed variables by v_i and the jth latent variables as l_j. The first block consists of a simple binary mapping designed to determine if the network can extract the correct relationships with no sources of noise. These relationships are written as

$$v_1 = l_1 \quad \text{\# a one to one map} \tag{26}$$

$$[v_2, v_3, v_4, v_5] = l_2 \quad \text{\# a one to many map} \tag{27}$$

$$[v_6, v_8] = l_3 \quad \text{\# one to many map that shares } v_8 \text{ with } l_4 \text{ mapping} \tag{28}$$

$$[v_7, v_8] = l_4 \quad \text{\# one to many map that shares } v_8 \text{ with } l_3 \text{ mapping} \tag{29}$$

$$v_9 = l_5 \quad \text{\# one to one map that is inverse of } v_1 0 \text{ (i.e. } 1 - v_{10}\text{)} \tag{30}$$

$$v_{10} = 1 - v_9 \quad \text{\# one to one map that is inverse of } v_9 \text{ (i.e. } 1 - v_9\text{)} \tag{31}$$

From this we model several different types of unambiguous relationships as described on the right, including logical relationships AND (v_2, v_3, v_4, v_5), OR (v_8) and NOT (v_{10}).

We build the next five blocks on exactly the same relationships. Block 2 utilises introduces noise into the mapping, requiring a uniformly sampled random value Unif$([0, 1])$ to be greater than a threshold $t = 0.25$ as well as the latent variable being equal to 1 for the relationship to be passed through to the observed variable(s). Unif$([0, 1]) > t$ returns 1 if the condition if fulfilled and 0 otherwise. Block 2 mappings can be written thus:

$$v_{11} = l_6 * (\text{Unif}([0, 1]) > t) \tag{32}$$

$$[v_{12}, v_{13}, v_{14}, v_{15}] = l_7 * (\text{Unif}([0, 1]) > t) \tag{33}$$

$$v_{16} = l_8 * (\text{Unif}([0, 1]) > t) \tag{34}$$

$$v_{17} = l_9 * (\text{Unif}([0, 1]) > t) \tag{35}$$

$$v_{18} = (l_8 = 1 \text{ OR } l_9 = 1) * (\text{Unif}([0, 1]) > t) \tag{36}$$

$$v_{19} = l_{10} = 1 * (\text{Unif}([0, 1]) > t) \tag{37}$$

$$v_{20} = 1 - v_{19} \tag{38}$$

Note that where the latent variable maps to many observed variables, a random value is sampled separately for each observed value so they are correlated rather than identical.

Block 3 reflects binary to continuous value [0,1] relationships in which the observed value(s) are zero if the latent value is zero but take a uniformly distributed random value [0,1] if the latent variable is 1. The mappings therein can be written as:

$$
\begin{align}
v_{21} &= l_{11} * \mathrm{Unif}([0,1]) \tag{39}\\
[v_{22}, v_{23}, v_{24}, v_{25}] &= l_{12} * \mathrm{Unif}([0,1]) \tag{40}\\
v_{26} &= l_{13} * \mathrm{Unif}([0,1]) \tag{41}\\
v_{27} &= l_{14} * \mathrm{Unif}([0,1]) \tag{42}\\
v_{28} &= (l_{13} \text{ OR } l_{14}) * \mathrm{Unif}([0,1]) \tag{43}\\
v_{29} &= l_{15} * \mathrm{Unif}([0,1]) \tag{44}\\
v_{30} &= 1 - v_{29} \tag{45}
\end{align}
$$

Block 4 introduces sampling from a parametric probability distribution, namely the beta distribution $\mathrm{Beta}(a, b)$, to reflect proportions and other continuous values bounded by 0 and 1. Here we aim to model situations where there are generally lower or higher values depending on the status of the latent variable. Therefore the main difference between this block and previous ones is that the observed value is sampled from one of two differently parameterised beta distributions depending if the latent variable is 1 of 0. The mappings are:

$$
\begin{align}
v_{31} &= l_{16} * \mathrm{Beta}(5,1) + (1 - l_{16}) * \mathrm{Beta}(1,5) \tag{46}\\
[v_{32}, v_{33}, v_{34}, v_{35}] &= l_{17} * \mathrm{Beta}(5,1) + (1 - l_{17}) * \mathrm{Beta}(1,5) \tag{47}\\
v_{36} &= l_{18} * \mathrm{Beta}(5,1) + (1 - l_{18}) * \mathrm{Beta}(1,5) \tag{48}\\
v_{37} &= l_{19} * \mathrm{Beta}(5,1) + (1 - l_{19}) * \mathrm{Beta}(1,5) \tag{49}\\
v_{38} &= (l_{18} \text{ OR } l_1 9) * \mathrm{Beta}(5,1) + (1\text{–}l_{18} OR l_{19}) * \mathrm{Beta}(1,5) \tag{50}\\
v_{39} &= l_{20} * \mathrm{Beta}(5,1) + (1 - l_{20}) * \mathrm{Beta}(1,5) \tag{51}\\
v_{40} &= 1 - v_{39} \tag{52}
\end{align}
$$

Block 5 is similar to block 3, where the observed value(s) are zero if the latent feature is zero. However, when the latent feature is 1 then the observed value is sampled from a Poisson distribution to simulate count data. All observed variables are rescaled by the maximum of $\tilde{v}_i = v_i/max(\mathbf{v}_i)$, but we leave this step out of the mapping equations below for simplicity and to aid comparison with the other blocks. We therefore write the mappings as:

$$
\begin{aligned}
v_{41} &= l_{21} * \text{Poiss}(10) && (53)\\
[v_{42}, v_{43}, v_{44}, v_{45}] &= l_{22} * \text{Poiss}(10) && (54)\\
v_{46} &= l_{23} * \text{Poiss}(10) && (55)\\
v_{47} &= l_{24} * \text{Poiss}(10) && (56)\\
v_{48} &= (l_{23} \text{ OR } l_{24}) * \text{Poiss}(10) && (57)\\
v_{49} &= l_{25} * \text{Poiss}(10) && (58)\\
v_{50} &= max(\mathbf{v}_{49}) - v_{49} && (59)
\end{aligned}
$$

In Block 6 we aim to model the situation where a perturbation to a cellular process elicits different levels of counts (such as gene expression in mRNA data for instance). The format is similar to block 4 but with Poisson distributions used instead of beta distributions. We adopt the same rescaling scheme as used in block 5.

$$
\begin{aligned}
v_{51} &= l_{26} * \text{Poiss}(30) + (1 - l_{26}) * \text{Poiss}(10) && (60)\\
[v_{52}, v_{53}, v_{54}, v_{55}] &= l_{27} * \text{Poiss}(30) + (1 - l_{27}) * \text{Poiss}(10) && (61)\\
v_{56} &= l_{28} * \text{Poiss}(30) + (1 - l_{28}) * \text{Poiss}(10) && (62)\\
v_{57} &= l_{29} * \text{Poiss}(30) + (1 - l_{29}) * \text{Poiss}(10) && (63)\\
v_{58} &= (l_{28} \text{ OR } l_{29}) * \text{Poiss}(30) + (1 - l_{28} \text{OR} l_{29}) * \text{Poiss}(10) && (64)\\
v_{59} &= l_{30} * \text{Poiss}(30)) + (1 - l_{30}) * \text{Poiss}(10) && (65)\\
v_{60} &= max(\mathbf{v}_{59}) - v_{59} && (66)
\end{aligned}
$$

Finally we include latent variables that are mapped to generating functions of more than one of the types in the blocks above. Simulated count data is again rescaled as before. The maps are given as

$$
\begin{aligned}
[v_{61}, v_{62}] &= l_{31} * [1, (\text{Unif}([0,1]) > t)] - (1 - l_{31}) * [0, 0] && (67)\\
[v_{63}, v_{64}] &= l_{32} * [\text{Unif}([0,1]), \text{Beta}(5,1)] + (1 - l_{32}) * [0, \text{Beta}(1,5)] && (68)\\
[v_{65}, v_{66}] &= l_{33} * [\text{Poiss}(10), \text{Poiss}(30)] + (1 - l_{33}) * [0, \text{Poiss}(10)] && (69)\\
[v_{67}, v_{68}, v_{69}, v_{70}, v_{71}, v_{72}] &= l_{34} * [1, (Unif([0,1]) > t), \text{Unif}([0,1]), \text{Beta}(1,5), \text{Poiss}(10), \text{Poiss}(30)]\\
&\quad + (1 - l_{34}) * [0, 0, 0, \text{Beta}(5,1), 0, \text{Poiss}(10)] && (70)
\end{aligned}
$$

RBM results on synthetic data

We trained 2000 networks using 80% of the synthetic data as the training set (chosen uniformly at random). The remainder of the data was used as a validation set for early stopping using the procedure described above. To investigate if overfitting occurs and if our early stopping procedure could identify potential overfitting, we allowed training to continue if overfitting was suspected and monitored the behaviour of the free energy. We found that overfitting was suspected in 123/2000 (6.15%) of the training runs and early stopping would have been invoked in these cases. We investigated what occurred by plotting the free energy of the training sets along with that of the validation set. We provide an example of a well-behaved profile (Figure S11A) and some examples of training runs that would have been stopped in Figures S11B-D. We observe that the free energy values oscillate with the periodicity of the cyclical learning rate described above and in the second half of training there are iterations where the free energy values decrease significantly across all sets, which corresponds to removal of dead hidden units in network pruning. In the samples where overfitting was not suspected, the free energy of the validation set decreased at the same rate at the free energy values of the training sets, remaining below the maximum value set by the training sets. This is exactly how we would expect the free energies to behave if there were no overfitting.

In training runs where overfitting was suspected, we found that this was generally because the free energy of the validation set had increased to a slightly higher value than the maximum value of the training sets; this always occurred during the latter half of training when network pruning was taking place (as in Figures S11B-C). This could indicate the network is starting to overfit. Occasionally there would be runs where the free energy of the validation set was notably higher than the maximum of the training sets (e.g. Figure S11D), which is more concerning and could indicate a higher degree of overfitting. However in every case we noted that in the general trend of the free energy of the validation set was to decrease until training stopped at the maximum number of iterations; this actually indicates that the data is not being overfit as we would expect the free energy to increase if the model was losing generalisability by incorporating aspects only present in the training sets. Therefore it is inconclusive whether these runs are actually overfit. Nonetheless, our early stopping procedure is very conservative and would have caught and removed these suspect runs, leaving only indisputably non-overfit runs, and so we are confident that overfit networks will not contribute to the final results.

We next performed training with early stopping enforced to investigate how well the network captures known relationships in the data. When training was complete, we amalgamated the weight matrices using the procedure described in Figure S10 and the final input feature encoding is given in Figure S12.

We can use the direct binary mappings of Block 1 to investigate how the RBM attempts to encode the relationships provided in the hidden data. The algorithm unambiguously identifies the one-to-one mapping of feature 1 to input 1 of this block, as well as the one-to-many mapping from feature 2 to inputs 2, 3, 4 and 5. In features 3 and 4, the algorithm can identify that input 6 and 8 arise from the same feature, as do inputs 7 and 8, but inputs 6 and 7 are not directly associated. The algorithm cannot identify the inverse relationship between inputs 9 and 10 and encodes them in two separate features (5 and 6). This is consistent with the way our approach is constructed as a positive weight matrix can only encode relationships in which a feature is active leading to inputs that are active rather than inactive feature giving rise to active inputs.

A similar pattern is seen in blocks 2-6, where the algorithm is generally able to extract the correct (or logical) associations encapsulated in the features. There are two exceptions to this, which we have highlighted by annotations A and B. Annotation A shows a different encoding of the situation in which the original features both map to the same input (as in features 3 and 4 in block 1). Here, these are encoded as three inferred features where the first two encode a strong association with each input exclusive to each feature and a weak association with the shared input and the third feature encodes a strong association with the shared input with weak associations to both of the exclusive inputs. It is important to note that although this encoding does not precisely replicate the original input mapping, the network has still learned a logical way of encoding this relationship. Therefore, we do not consider this encoding incorrect. Annotation B shows inputs that are not assigned to any feature. Both of these are one-to-one mappings in block 6 (Poisson low/high) indicating that these relationships in this data type might be too subtle for the algorithm to distinguish. Note the one-to-one mapping of the 9th input in the block and the one-to-many mapping are identified correctly. In block 7, the block consisting of one-to-many mappings of data of multiple types, the algorithm identifies the correct number of features (4) and the correct inputs are assigned to each feature. These results provide empirical evidence that our RBM algorithm can extract relationships across a number of data types.

Consistency of results

There are several sources of variation between runs on the same data set: the RBM is intrinsically a probabilistic network, training it is a stochastic process and we are using different sets of data in each run. However, although variation between the features extracted in each run is inevitable, it is important that they are consistent as an ensemble so we can be confident that the result is stable and the final amalgamated weights reflect the true relationships in the data.

To quantify the consistency between feature sets, we use an approach based on the Hamming distance. If we describe a latent feature as a binary string that is equal to 1 with a non-zero weight is present and zero elsewhere, we can then calculate the Hamming distance between individual features, which returns the number of input mappings, m, that are not shared between those features. We can use this to identify which feature in set $\mathbf{F}_j$ is *equivalent* to a given feature in set $\mathbf{F}_i$ (as that has the minimum Hamming distance) and then identify the feature in set F_i that is most *incongruous* to their equivalent feature (as this has the maximum Hamming distance). We write the Hamming distance of the incongruous feature as $\tilde{m}_{i,j}$, which can be expressed as

$$\tilde{m}_{i,j} = \max_i \min_j (D_{hamming}(\mathbf{F}_i, \mathbf{F}_j)) \tag{71}$$

We can use this metric to assess how the input/feature mapping differs between runs. We randomly sampled $w = 1000$ weight matrices (without replacement) from the 2000 matrices generated by networks trained on the synthetic data and amalgamated these as described above. We repeated this 100 times to give 100 feature sets, and then we calculated $\tilde{m}_{i,j}$ for all pairwise combinations of i and j. Of the 10,000 resulting comparisons, the greatest number of differences between equivalent features was 2, which occurred 72 times (0.72% of the time). Figure S13 shows the histogram of the $\tilde{m}_{i,j}$ values for $w = 1000$, which reveals that the most common difference was 1, which occurred 6001 times (60.01%). There was no difference in the feature maps in 39.27% of the runs. This is remarkably consistent given the aforementioned sources of variation.

Network training on real data

We used the real data to train various versions of the networks across a number of runs to obtain distinct outputs for use in the analysis. The goal of the first run was to extract the amalgamated weight matrix that describes the input to feature mapping (Figure S2). We used this to determine the latent feature representation used in the clustering (MP Figure 1) by training a new network run with the weight matrix initialised to the amalgamated weight matrix and setting the weight learning rate to zero. Learning of the biases was enabled, as these may be different to the biases in the previous networks due to the removal of low magnitude weights. Once the remaining network parameters have converged during training, taking further iterations is equivalent to sampling the hidden units/feature representation for each patient. We therefore averaged the hidden unit values taken every 10 iterations during the final 1000 iterations to obtain the final feature representation.

The input-feature map was used to extract an informed subset of genetic alterations from the original inputs - these were used to determine associations in the ARBS analysis (MP Figure 2), as well as the inputs for the Ordering Analysis (MP Figure 3).

Two-stage clustering

The dimensionality of the feature representation is still quite large for conventional clustering techniques. Therefore we adopted a two-stage approach where we first clustered by those features that were most informative of clinical outcome, calculated the centroids of these first-stage clusters for all features, and then clustered these in the second-stage of clustering to produce MP Figure 1. Here we provide more details on identification of informative features using a discrimination score and the clustering methods used.

Discrimination score

There have been several methods proposed for quantifying the relative importance of the units of a neural network[60]. However, most of these are generally formulated to discover the inputs that are important in discerning the output[61,62]. In our application, we wish to quantify the discriminative capacity of each of the features (hidden layer) with respect to the clinical outcome. As we utilise non-negative weights to determine the relevance of the inputs to the hidden units in the feature extraction, for consistency we adopt a similar strategy to determine the relevance of the hidden units to adverse clinical outcome as determined by biochemical relapse.

To obtain the discrimination scores for each feature, we modified the architecture of the base RBM so that it was similar to ClassRBM[63]. This adds an extra classification layer, which is fully connected to the hidden layer, the units of which contain the values of the classes. In ClassRBM, there is another set of weights that denote the strength of the connection between the hidden and classification layers, and these are trained in the same bi-directional fashion as the input weights.

However, in our application we wish to uncover underlying relationships in the data (encapsulated by the features) in an unbiased way and then determine how relevant these features are to determining the clinical outcome. We therefore performed the learning of the latent feature representation and the discrimination scores separately to ensure that learning the classification weights to ensure that the latent representation remains unbiased by the knowledge of the clinical outcome, and the algorithm for feature learning described above can still be considered as unsupervised. This was done by fixing the input weights to the amalgamated weight matrix described above but then training the class weights using contrastive divergence as described above.

We also enforced a non-negative constraint on these class weights, similar to the input weights. To get our *discrimination score*, we take the absolute value of the weights corresponding to relapse minus the weights corresponding to no-relapse, $\mathbf{s}$. This can be expressed mathematically as

$$\mathbf{s} = |\mathbf{c}_r - \mathbf{c}_{r'}|, \tag{72}$$

were $\mathbf{c}_r$ are the class-weights associated with relapse, and $\mathbf{c}_{r'}$ are those associated with no relapse.

These $\mathbf{s}$ values can be considered as heuristic quantity relating importance of the corresponding feature to the clinical output, similar to how the component loadings quantify the explained variance of the corresponding principal component in principal component analysis (PCA). As there is no set rule for determining the number of features, so we followed a similar approach to that conventionally used in PCA and selected the number of features using the cumulative distribution. We chose a cut off of 0.9 of the total cumulative discrimination score, which resulted in 14 out of 30 features being selected for the initial clustering phase.

Clustering

Clustering of tumours was performed on the latent feature representation in a two-stage process to facilitate the identification of clusters that were relevant to clinical outcome. As the feature representation for each patient can be considered as a vector containing the probabilities that the corresponding feature is active, it is appropriate to use a distance measure that quantifies the distance between probabilities. As such, we calculated the mean Jensen-Shannon (J-S) divergence[64] between tumours in a pairwise fashion.

For a pair of patients A and B, represented by the latent feature representation in hidden layers $\mathbf{h}_A$ and $\mathbf{h}_B$, the mean J-S divergence can be written as

$$JSD(\mathbf{h}_A \parallel \mathbf{h}_B) = \frac{1}{2K} \sum_{i=1}^{K} \left[h_{A,i} \log\left(\frac{h_{A,i}}{m_i}\right) + h_{B,i} \log\left(\frac{h_{B,i}}{m_i}\right) \right], \qquad (73)$$

where $\mathbf{m} = \frac{1}{2}(\mathbf{h}_A + \mathbf{h}_B)$, is the midpoint of $\mathbf{h}_A$ and $\mathbf{h}_B$. The additive terms in the square brackets in Equation 73 represent the Kullback-Leibler divergence between each element of the latent feature representation for either patient and the corresponding element of the midpoint vector, $\mathbf{m}$.

As we are not using a Euclidean distance metric, clustering through k-means is not appropriate and so we used k-medoid clustering for the first stage; this is similar to k-means but selects a representative data point (*medoid*) as the centroid for each cluster instead of the mean. Using the silhouette method[65], we determined that 11 clusters was optimal. For the second stage of clustering, we used hierarchical clustering to cluster the medoids themselves (again using the J-S divergence), and this was used to generate and order clusters by the dendrogram MP Figure 1.

DNA breakpoint proximity to androgen receptor binding site

To examine the proximity of DNA breakpoints to androgen receptor binding sites (ARBS), we designed a permutation approach that quantifies the departure from a random distribution of the breakpoints across the genome. We downloaded processed ChIP-seq data targeting AR for 13 primary prostate cancer tumours from Gene Expression Omnibus (accession GSE70079)[66] and amalgamated them for use as the ARBS locations. For each of our 159 samples, we simulated the scenario whereby breakpoints were randomly distributed across the genome. The simulation of breakpoints was performed chromosome-wise. For each chromosome we simulate 1000 sets of N breakpoint positions, where N is the number of breakpoints we observed on that chromosome. These positions are randomly distributed (with a uniform distribution) across the full chromosome. Therefore the simulations intrinsically take into account the size of the chromosomes. We used the R package RegioneR[67] with genome assembly GRCh37, masked for assembly gaps (AGAPS mask) and intra-contig ambiguities (AMB mask) to keep the possible chromosomal locations consistent with what could be observed in the real data.

To detect significant departure from a uniform random distribution, we calculated the proportion of breakpoints within 20,000 base pairs (bp) of an ARBS for the observed and permuted data (B_{obs} and B_{perm}, respectively). If $B_{obs} > p_{97.5\%}(B_{perm})$, the tumour was classified as Enriched, else if $B_{obs} < p_{2.5\%}(B_{perm})$, the tumour was classified as Depleted. Otherwise the difference is not significant and the tumour was classified as Indeterminate. The level of enrichment or depletion of breakpoints in the proximity of ARBS used in MP Figure 2A was estimated according to the following formula:

$$D = B_{obs} - \tilde{B}_{perm}. \qquad (74)$$

To check the method was not inherently biased, we performed the analysis on the breakpoints derived from the UK data set (as in MP Figure 2) and compared these to the classes derived if the position of the AR binding sites was distributed uniformly across the genome (Figure S14). As expected, we find that almost all tumours in the randomised set were classed as Indeterminate, in stark contrast to the real data.

The agglomerative hierarchical clustering of the ARBS groups across Australian, Canadian and UK data sets was generated using the R package `pvclust`[68] v2.0.0 using the ward.D2 clustering method with squared Euclidean distance (100,000 iterations). This package also enabled the estimation of the Approximately Unbiased Multiscale Bootstrap (AU) p-values for the Depleted group. These clustering results were confirmed by a partitional clustering approach using the R packages `cluster` v2.1.0 and `factoextra` v1.0.5.

ARBS pairs required for DNA loop formation

In MP Figure 5 we investigated the role of AR in the formation of DNA loops that can precipitate DNA double strand breaks (DSBs). For each ARBS that was previously identified as proximal to a DNA breakpoint in each patient, we determined the proportion that were also proximal to another ARBS, as required by the mechanism for DNA loop formation described previously[69] and depicted in MP Figure 5A. As it has been shown that the two ARBS involved in the DNA loop involved in the *TMPRSS2/ERG* fusion are separated by 19972 base pairs[69], we set the threshold for the second proximal ARBS as 30,000 base pairs in the direction away from the DNA breakpoint. p-values were determined as before.

In all box plots the red line denotes the median, the blue box encapsulates the interquartile range, and the black dashed lines denote the range of data not considered outliers; outliers (red dots) are as defined in the Matlab boxplot() function default settings. The size of the angular 'notch' corresponds to a confidence interval around the median, such that if two notches are not overlapping then there is approximately 95% confidence that the median of the two groups differ. The folded notch in MP Figure 5D indicates that the notch extends past the interquartile range by the folded amount.

Ordering

We previously estimated consensus ordering of events by estimating phylogenetic trees from the cancer cell fraction (CCF) that contained each aberration, and applying the Bradley-Terry model to determine the most consistent order of events[11]. We recently released a study[70] that improves this approach using a Plackett-Luce model[71,72], which we also utilise in this study. We provide a complete description of the method for reproducibility.

There are a number of sources of uncertainty when attempting to determine the order of events from bulk DNA sequencing. In particular, we often cannot infer the true phylogenetic tree for each patient, and furthermore it is impossible to determine the relative timing of events on parallel branches. However, we can estimate the set of possible trees using the relative cancer cell fractions (CCFs) of the genomic aberrations involved, and from these we can estimate a set of possible orderings. Therefore we created an algorithm where we (*sampled*) a single possible tree from the data, and using this we sampled a viable order of events for each patient. This is repeated multiple times so that the uncertainty in these estimates is encapsulated in the output distributions. Algorithms of this type are called *Monte-Carlo* simulations to emphasise the use of randomness in the procedure.

We adopted the Plackett-Luce model[71,72] to construct a probability distribution over the relative rankings of a finite set of items, the parameters of which can then be estimated from a number of individual rankings. This can be used to quantify the expected rank of each item relative to the others across the population. In our application, an item corresponds to an event, namely the emergence and fixation of a novel copy number alteration (CNA) identified in the extracted features. Ranking these events therefore relates to the order in which they would be expected to occur. We also utilised a Plackett-Luce mixture model[20], which allow us to determine whether there are subpopulations in the data with different orderings.

The Plackett-Luce model

Given a set of CNA occurrences for each patient with associated subclonality, we would like to infer the order which these events generally occur. To do this we used a Plackett-Luce model, which is formulated as a *ranking* method, and returns a value quantifying the ranking preference. We use a different interpretation, namely the *ordering*, which is defined as the inverse of the ranking preference[73]. Like the Bradley-Terry model, the Plackett-Luce model does not return any temporal information outside the expected order of events.

We have a set of N copy number events we are interested in,

$$C = \{c_1, c_2, \ldots, c_N\}, \tag{75}$$

then we can apply Luce's choice axiom[71], which states that the probability of selecting one event over another from a set of events is independent of the presence or absence of the other events in the set. We can therefore write the probability of observing event i as

$$P(c_i|C) = \frac{\alpha_i}{\sum_j \alpha_j}, \tag{76}$$

where $\{\alpha_i\}$ are the coefficients that quantify the relative probability of observing the i^{th} event. To reflect the ordering aspect of our application we refer to this value as the *proclivity.* Plackett[72] used this formalism to construct a generative model in which all N events are randomly sampled from C without replacement (i.e. a *permutation*). If we let Λ correspond to a permutation of the set C such that $\lambda_k \in C$ and $\lambda_1 \prec \lambda_2 \prec \ldots \prec \lambda_N$, then we write the probability density of a single ordering as

$$P(\Lambda) = \prod_{k}^{N} \frac{\alpha_{\lambda_k}}{\sum_{j \in \Lambda^{(k)}} \alpha_j}, \tag{77}$$

where α_{λ_k} is the proclivity associated with event λ_k, and $\Lambda^{(k)} = \{\lambda_k, \lambda_{k+1}, \ldots, \lambda_N\}$ is the set of possible events after $k-1$ events have occurred.

Plackett-Luce mixtures

We hypothesised that there may be more than one set of copy number orderings present in our population, and so analysing all events in one ordering scheme may not be appropriate. Furthermore, the inhibition of AR-associated breakpoints implies that some CNAs may be found more frequently with a select set of others, which is in violation of Luce's choice axiom. We therefore implemented a mixture modelling approach[20,73], which reinstates Luce's choice axiom as the selection of each CNA can be considered as independent conditional on the mixture component. Such a finite mixture model assumes that the population consists of a number, G, of subpopulations. In this setting the probability of observing the ordering Λ_s for the s^{th} sample is

$$P(\Lambda_s) = \sum_{g}^{G} \omega_g P_g(\Lambda_s), \tag{78}$$

where ω_g are the weight parameters (not to be confused with the weight matrix in the RBM) that quantify the probability that sample s belongs to subgroup g. The appropriate parameter values can be determined using maximum likelihood estimation via an EM algorithm[20]. The number of mixture components can be chosen using the Bayesian Information Criterion (BIC) estimation, which is given by

$$BIC = N \log(M) - 2\ell(\Theta_{ML}), \tag{79}$$

where Θ_{ML} is the parameter set that maximises the log-likelihood $\ell(\cdot)$, N is the number of parameters, and M is the number of samples.

Implementation

The general formulation of the Plackett-Luce model takes a matrix containing the sequence of events for each patient as its input. However, we do not know the order in which these events occurred, only the presence and cancer cell fraction (CCF) of each CNA for each patient. As such, we first estimate the phylogenetic trees for each patient, and then determine the order of events from this. As we only have one tissue sample for each patient, there is often uncertainty in the tree topology and the possible sequence of events, and so we use a Monte-Carlo sampling scheme in which we sample the trees and sequence of events, and use these to estimate the distribution of possible orderings through the Plackett-Luce model. Samples with 0 or 1 CNA were not used in this analysis.

Another issue arises due to censoring, which occurs when the sample is taken before all aberrations that would occur have occurred, resulting in missing data. These are called partial-orderings in the Plackett-Luce framework, and the general approach to addressing this is to reformulate the model so that all missing events are implicitly ranked lower than the observed data[20,74]. This may not be appropriate for our analysis as we may have multiple subgroups, and we anticipate that distinct aberrations may have similar or equivalent effects in each subtype and thus will rarely co-occur despite being indicative of the same type. For instance, the absence of a very early aberration may be due to the occurrence of another less frequent aberration, so including it at the bottom of the order would bias the rankings toward more frequent aberrations. As such, our algorithm works in two phases:

1. Determine the number of mixture components and assign patients to each component,
2. Estimate the ordering profiles of each component.

These are distinct as we treat the creation of the phylogenetic trees in a slightly different way in each of these processes to account for censoring. When estimating the number of components, we calculate trees only using the observed CNAs. However, when estimating the full ordering profiles, we introduce another sampling step into our Monte-Carlo scheme where we explicitly sample a number of additional CNAs with probability proportional to the subclonality of the aberration in tumours of each mixture component. Sampling in this way reduces the bias toward more frequent aberrations.

Assign samples to mixture components

In the first phase, we

1. Sample phylogenetic trees for each patient,
2. Sample sequence of events for each patient that are consistent with trees,
3. Calculate Bayesian Information Criterion (BIC) for 1-10 mixture components,
4. Repeat steps 1-3 1000 times,
5. Determine number of mixture components which consistently had lowest BIC score,
6. Assign patients to mixture components.

The phylogenetic trees are created by initially sorting the CNAs of each patient in descending order of CCF obtained from the output of the Battenberg algorithm, iterating through them and sampling the possible parents with uniform probability. The CCF of a parent cannot be greater than the sum of the CCF of their children, so viable parents are defined as ones where their CCF is greater than that of their current children plus the CCF of the CNA under consideration. The position in the sequence when the CNA occurred is sampled as any position after the parent, with uniform probability. The ordering estimates and assignment to the mixture components using the R package `PLMIX` as this incorporates mixture models and partial rankings (so the absence of a CNA from a sequence would not penalise its position in the ordering). A vector of assignments was retained for each sample run, and the final assignment was determined by the most frequent assignment over the course of 1000 runs.

Estimate ordering profiles of each component

In the second phase, we

1. Sample phylogenetic trees for each patient,
2. Sample sequence of events for each patient that are consistent with trees,
3. Augment sequence with additional CNAs to alleviate censorship bias,
4. Calculate ordering profiles for each mixture component,
5. Repeat steps 1-4 1000 times,
6. Amalgamate results to determine final ordering profiles of each mixture component.

The phylogenetic trees and sequence of events were initially determined as before. However, instead of utilising partial rankings in the PL model, we explicitly augmented the data with additional CNAs to account for those unobserved due to censorship. The probability of CNA being added to the sequence of events is equal to the proportion of subclonal occurrences relative to the total number of occurrences in the subpopulation defined by the mixture component. This can be written as

$$P(\tilde{c}_{ig}) = \frac{N_{sub}(c_{ig})}{N_{tot}(c_{ig})}, \tag{80}$$

where $N_{sub}(\cdot)$ and $N_{total}(\cdot)$ denote the number of subclonal and total occurrences respectively of CNA c_i in mixture component g. As events that are predominantly subclonal have a higher chance of being unobserved due to censorship, this sampling scheme will mitigate this to a degree. Conversely, events that are predominantly clonal (i.e. early) may be unobserved due to factors other than censoring, and these have a reduced chance of being imputed. Calculating these values using the patient samples for each mixture components rather than the entire population means that only CNA subclonality relevant to each subpopulation are considered. Imputation is performed by drawing a uniform random number, r, for each patient and including the CNA in the set of additional CNAs for each patient if $P(\tilde{c}_{ig}) < r$. The set of additional CNAs for each patient are shuffled uniformly and added to the sequence. We then calculate the ordering for each mixture component individually using the Plackett-Luce model without partial ranking. This process is repeated 1000 times and the proclivity for each CNA is calculated and used to create an empirical distribution for proclivity for each CNA, which are used to create the box-plots in MP Figure 3.

Synthetic data

We generated a synthetic data set to evaluate our method. We simulated two subpopulations, A and B that each had exclusive sets of 5 'early' and 5 'late' CNAs as well as common sets of 5 early and 5 late CNAs. These can be written as

$$
\begin{aligned}
E_A &= \{c_1, c_2, c_3, c_4, c_5\}, && (81)\\
E_B &= \{c_6, c_7, c_8, c_9, c_{10}\}, && (82)\\
L_A &= \{c_{11}, c_{12}, c_{13}, c_{14}, c_{15}\}, && (83)\\
L_B &= \{c_{16}, c_{17}, c_{18}, c_{19}, c_{20}\}, && (84)\\
E_C &= \{c_{21}, c_{22}, c_{23}, c_{24}, c_{25}\}, && (85)\\
L_C &= \{c_{26}, c_{27}, c_{28}, c_{29}, c_{30}\}. && (86)
\end{aligned}
$$

We used these sets to simulate the order of the events in each tumour, and then created a set of CCF values consistent with this order. The main principle in the simulation is that early events will generally occur before late events, but there is no intrinsic order in the sets of early and late events themselves.

To obtain the set of events for each simulation, Y, we first sampled how many events occurred by the time of sampling from a Poisson distribution, $e \sim Poiss(10)$ followed by the subpopulation S to draw from, with $p(S = A) = p(S = B) = 0.5$. To reflect the early/late ordering and exclusive/common nature of the CNAs, we first sampled 5 events (or e events if $e < 5$) from the pooled set of common and early events for that subpopulation, $P_E = \{E_S \cup E_C\}$, without replacement, that is $Y_E \subseteq P_E, |Y_E| = 5$. If $e > 5$, we then pooled the events that had not been sampled from P_E already, denoted here as P'_E, with the set of late events $P_L = \{L_S \cup L_C\}$ and sampled the remaining $e - 5$ events from this set, $Y_L \subseteq \{P'_E \cup P_L\}$. We then randomly sampled the order in which the events in Y_E occurred, followed by those in Y_L. We then sampled how many of these events were clonal, e_c, uniformly at random as assigned these a CCF of 1. To obtain the CCFs values of the subclonal population we ranked the subclonal events, $r_s \in \{1, 2, \ldots, e_s\}$, and assigned a CCF value as a linear function of their rank $CCF_k = 1 - r_s/(e_s + 1)$.

We created synthetic CCF values for 200 tumours and used these as inputs into our algorithm described above. The BIC scores are shown in Figure S15A, where two mixture components has the lowest BIC score and so the algorithm has identified the correct number of subpopulations. These subpopulations were used in to establish ordering profiles, which are shown in Figure S15B-C. We find that the algorithm has correctly identified all of the unique early and late events for each subpopulation, as well as the common early and late events.

BIC scores for real data

Bayesian Information Criterion (BIC) scores were determined for each mixture component for each of the 1000 runs are shown in Figure S16. The BIC score was lowest for two mixture components for every sampled ordering, and so this was taken as the value to use in subsequent analysis.

Statistical model of evotype convergence

We created a statistical model describing how the probability of convergence to the Canonical or Alternative evotypes changes as genetic alterations accumulate (MP Figure 4D). We assume that the accumulation of such aberrations in each individual tumour followed a stochastic process in which the order and relative timing of the aberrations occurred with some degree of randomness/stochasticity. Similar to the Ordering analysis, we utilised a statistical algorithm in which we simulated a number of possible aberrations consistent with the possible phylogenetic trees, and then estimated the probability that tumours with these aberrations converged to the Canonical-evotype (the probability of convergence to the Alternative-evotype is 1 minus the probability of convergence to the Canonical-evotype). As many of the genetic alterations will occur clonally (i.e. with $CCF = 1$) there is considerable uncertainty in the order in which they would have occurred. We incorporate this uncertainty into our approach using a method akin to *propagation of errors through Monte-Carlo simulation*[75], in which we sample (with uniform probability) from the space of possible trees for each tumour and calculate the probability of tumours with those genetic alterations converging to the Canonical-evotype. Repeating this random sampling many times provides a distribution of outputs that incorporates the uncertainty arising from our inputs in a principled fashion, enabling downstream analysis.

Our algorithm is outlined in Figure S17. The algorithm iterates through an increasing number of aberrations (Loop i), performing several Monte-Carlo repeats of ordering samples (Loop j).

As individual evolutionary trajectories involve the stochastic accumulation of multiple genomic aberrations, is it impossible to specify each evolutionary route. However, we can determine common *modes* of evolution by tracking the genetic alterations prevalent in tumours at the point of convergence to either evotype in our model. Through this we can identify paths in the probability density surface plot that correspond to the accumulation of these genetic alterations (black dashed lines, MP Figure 4D), and the aberrations that distinguish them from each other

Modelling the stochastic accumulation of genomic aberrations

We model the accumulation of aberrations in a tumour as a Poisson process[76]. For each iteration of Loop i we update the mean number of aberrations x_i, this is then used as the input parameter to a Poisson random number generator to draw the number of aberrations to be sampled, n, in each iteration of Loop j. We then identified those tumours with sufficient aberrations and selected one with uniform probability, and used the data for these to sample a phylogenetic tree using the relative CCFs of the aberrations. We then used the phylogenetic trees to sample an order of occurrence for the aberrations, and retained the first n. The aberrations used were the *SPOP* mutations and the CNAs identified in the feature extraction; inter-intra chromosomal breakpoints, ETS status and chromothripsis are not included as these do not have associated CCFs and therefore cannot be used to determine the order of events.

We again use the data to calculate probability that tumours with the set of aberrations will be assigned to the Canonical-evotype. For a set of sampled aberrations, $A_j = \{a_1, a_2, \ldots, a_n\}$, we identified the patients for which $A_j \subseteq P_k$, where P_k denotes the full set of aberrations present in patient k. We can then identify which of these were assigned to the Canonical-evotype. We can now calculate the probabilities

$$\begin{aligned} p(A_j) &= \frac{N(A_j \subseteq P_k)}{N(P_k)}, & (87) \\ p(Canonical \cap A_j) &= \frac{N(Canonical \cap (A_j \subseteq P_k))}{N(P_k)}, & (88) \end{aligned}$$

where $N(\cdot)$ denotes the number of tumours that obey the condition in brackets. We can now calculate the conditional probability

$$p(Canonical|A_j) = \frac{p(Canonical \cap A_j)}{p(A_j)}. \quad (89)$$

We performed 100,000 samples and thus obtained 100,000 values for each $p(Canonical|A_j)$. We input these values into a nonparametric density estimation scheme using Gaussian kernels with bandwidth 0.025. As we are estimating the probability density function of a set of probabilities, which are bound at $[0, 1]$, we ensured support only over this interval using the reflection method[77]. We performed this sampling step for $x_i \in \{0, 0.01, 0.02, \ldots, 10\}; i \in 1, 2, \ldots, 1000$.

Identifying genetic alterations in the convergent evolutionary trajectories

We used our model simulations to investigate the common evolutionary trajectories involved in convergence to each evotype (black dashed lines MP Figure 4D) as well as the aberrations that characterise them. In the modelling process, we recorded the order of genetic alterations for each of the trajectories used to calculate the pdf. We extracted each trajectory that had converged to the canonical or alternative evotypes (i.e. had a $p(Canonical|A_j) = 0$ or 1) and assigned these into sets by the number of genetic alterations in the trajectories i.e. $\{A_1\}, \{A_2\}, \ldots, \{A_{10}\}$. We than ran a filtering step for each set where we removed any trajectories that had occurred in sets corresponding to fewer genetic alterations, meaning we were left with trajectories that only converged to either evotype with the final genetic alteration for each set. We can then identify the position and frequency of occurrence of each genetic alteration in each set. The results of this are plotted in the bottom pane in Figures S7 and S8 for Canonical- and Alternative-evotype tumours respectively. Using this information we can calculate the pdf values for frequent combinations of genetic alterations in order, and use these to create the representative paths through the probability density (black dashed lines; MP Figure 4D).

CRUK ICGC Prostate Group members

Additional to the named authors, the CRUK ICGC Prostate Group also contains the following members:

- Adam Lambert, University of Oxford, Oxford, UK

- Anne Babbage, Hutchison/MRC Research Centre, Cambridge University, Cambridge, UK

- Clare L. Verrill, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

- Claudia Buhigas, Norwich Medical School, University of East Anglia, Norwich, UK

- Dan Berney, Department of Molecular Oncology, Barts Cancer Centre, London, UK

- Ian G. Mills, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

- Nening Dennis, Royal Marsden NHS Foundation Trust, London and Sutton, UK

- Sarah Thomas, Royal Marsden NHS Foundation Trust, London and Sutton, UK

- Sue Merson, The Institute Of Cancer Research, London, UK

- Thomas J. Mitchell, Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK

- Wing-Kit Leung, Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge, UK.

- Alastair D. Lamb, Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK

References

1. Nowell, P.C. (1976). The clonal evolution of tumor cell populations. Science *194*, 23–8.

2. Fearon, E.R., Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. Cell *61*, 759–67. doi:10.1016/0092-8674(90)90186-i.

3. Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Rosebrock, D., Mitchell, T.J., Rubanova, Y., Anur, P., Yu, K., et al. (2020). The evolutionary history of 2,658 cancers. Nature *578*, 122–128. doi:10.1038/s41586-019-1907-7.

4. Nangalia, J., Nice, F.L., Wedge, D.C., Godfrey, A.L., Grinfeld, J., Thakker, C., Massie, C.E., Baxter, J., Sewell, D., Silber, Y., et al. (2015). DNMT3A mutations occur early or late in patients with myeloproliferative neoplasms and mutation order influences phenotype. Haematologica *100*, e438–42. doi:10.3324/haematol.2015.129510.

5. Ortmann, C.A., Kent, D.G., Nangalia, J., Silber, Y., Wedge, D.C., Grinfeld, J., Baxter, E.J., Massie, C.E., Papaemmanuil, E., Menon, S., et al. (2015). Effect of mutation order on myeloproliferative neoplasms. N. Engl. J. Med. *372*, 601–612. doi:10.1056/NEJMoa1412098.

6. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic classification and prognosis in acute myeloid leukemia. N. Engl. J. Med. *374*, 2209–2221. doi:10.1056/NEJMoa1516192.

7. Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., Yoon, C.J., Ellis, P., Wedge, D.C., Pellagatti, A., et al. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. Blood *122*, 3616–27. doi:10.1182/blood-2013-08-518886.

8. Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., et al. (2005). Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science *310*, 644–8. doi:10.1126/science.1117679.

9. Kaffenberger, S.D., Barbieri, C.E. (2016). Molecular subtyping of prostate cancer. Curr. Opin. Urol. *26*, 213–8. doi:10.1097/MOU.0000000000000285.

10. Luca, B.A., Brewer, D.S., Edwards, D.R., Edwards, S., Whitaker, H.C., Merson, S., Dennis, N., Cooper, R.A., Hazell, S., Warren, A.Y., et al. (2018). Desnt: A poor prognosis category of human prostate cancer. Eur. Urol. Focus *4*, 842–850. doi:10.1016/j.euf.2017.01.016.

11. Wedge, D.C., Gundem, G., Mitchell, T., Woodcock, D.J., Martincorena, I., Ghori, M., Zamora, J., Butler, A., Whitaker, H., Kote-Jarai, Z., et al.

(2018). Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nat Genet *50*, 682–692. doi:10.1038/s41588-018-0086-z.

12. The Cancer Genome Atlas Research Network. (2015). The molecular taxonomy of primary prostate cancer. Cell *163*, 1011–25. doi:10.1016/j.cell.2015.10.025.

13. Fraser, M., Sabelnykova, V.Y., Yamaguchi, T.N., Heisler, L.E., Livingstone, J., Huang, V., Shiah, Y.J., Yousif, F., Lin, X., Masella, A.P., et al. (2017). Genomic hallmarks of localized, non-indolent prostate cancer. Nature *541*, 359–364. doi:10.1038/nature20788.

14. Boyd, L.K., Mao, X., Lu, Y.J. (2012). The complexity of prostate cancer: genomic alterations and heterogeneity. Nat. Rev. Urol. *9*, 652–64. doi:10.1038/nrurol.2012.185.

15. Gerhauser, C., Favero, F., Risch, T., Simon, R., Feuerbach, L., Assenov, Y., Heckmann, D., Sidiropoulos, N., Waszak, S.M., Hubschmann, D., et al. (2018). Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. Cancer Cell *34*, 996–1011 e8. doi:10.1016/j.ccell.2018.10.016.

16. Espiritu, S.M.G., Liu, L.Y., Rubanova, Y., Bhandari, V., Holgersen, E.M., Szyca, L.M., Fox, N.S., Chua, M.L.K., Yamaguchi, T.N., Heisler, L.E., et al. (2018). The evolutionary landscape of localized prostate cancers drives clinical aggression. Cell *173*, 1003–1013 e15. doi:10.1016/j.cell.2018.03.029.

17. Haffner, M.C., Aryee, M.J., Toubaji, A., Esopi, D.M., Albadine, R., Gurel, B., Isaacs, W.B., Bova, G.S., Liu, W., Xu, J., et al. (2010). Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. Nat. Genet. *42*, 668–75. doi:10.1038/ng.613.

18. Weischenfeldt, J., Simon, R., Feuerbach, L., Schlangen, K., Weichenhan, D., Minner, S., Wuttig, D., Warnatz, H.J., Stehr, H., Rausch, T., et al. (2013). Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. Cancer Cell *23*, 159–170. doi:10.1016/j.ccr.2013.01.002.

19. Augello, M.A., Liu, D., Deonarine, L.D., Robinson, B.D., Huang, D., Stelloo, S., Blattner, M., Doane, A.S., Wong, E.W.P., Chen, Y., et al. (2019). CHD1 loss alters AR binding at lineage-specific enhancers and modulates distinct transcriptional programs to drive prostate tumorigenesis. Cancer Cell *35*, 817–819. doi:10.1016/j.ccell.2019.04.012.

20. Mollica, C., Tardella, L. (2017). Bayesian Plackett-Luce mixture models for partially ranked data. Psychometrika *82*, 442–458. doi:10.1007/s11336-016-9530-0.

21. Metzger, E., Willmann, D., McMillan, J., Forne, I., Metzger, P., Gerhardt, S., Petroll, K., von Maessenhausen, A., Urban, S., Schott, A.K., et al. (2016). Assembly of methylated KDM1A and CHD1 drives androgen receptor-dependent transcription and translocation. Nat. Struct. Mol. Biol. *23*, 132–9. doi:10.1038/nsmb.3153.

22. Kluth, M., Runte, F., Barow, P., Omari, J., Abdelaziz, Z.M., Paustian, L., Steurer, S., Christina Tsourlakis, M., Fisch, M., Graefen, M., et al. (2015). Concurrent deletion of 16q23 and PTEN is an independent prognostic feature in prostate cancer. Int. J. Cancer *137*, 2354–63. doi:10.1002/ijc.29613.

23. Kluth, M., Harasimowicz, S., Burkhardt, L., Grupp, K., Krohn, A., Prien, K., Gjoni, J., Hass, T., Galal, R., Graefen, M., et al. (2014). Clinical significance of different types of p53 gene alteration in surgically treated prostate cancer. Int. J. Cancer *135*, 1369–80. doi:10.1002/ijc.28784.

24. Shenoy, T.R., Boysen, G., Wang, M.Y., Xu, Q.Z., Guo, W., Koh, F.M., Wang, C., Zhang, L.Z., Wang, Y., Gil, V., et al. (2017). CHD1 loss sensitizes prostate cancer to DNA damaging therapy by promoting error-prone double-strand break repair. Ann. Oncol. *28*, 1495–1507. doi:10.1093/annonc/mdx165.

25. Boysen, G., Rodrigues, D.N., Rescigno, P., Seed, G., Dolling, D., Riisnaes, R., Crespo, M., Zafeiriou, Z., Sumanasuriya, S., Bianchini, D., et al. (2018). SPOP-mutated/CHD1-deleted lethal prostate cancer and abiraterone sensitivity. Clin. Cancer. Res. *24*, 5585–5593. doi:10.1158/1078-0432.CCR-18-0937.

26. Rodrigues, L.U., Rider, L., Nieto, C., Romero, L., Karimpour-Fard, A., Loda, M., Lucia, M.S., Wu, M., Shi, L., Cimic, A., et al. (2015). Coordinate loss of MAP3K7 and CHD1 promotes aggressive prostate cancer. Cancer Res. *75*, 1021–34. doi:10.1158/0008-5472.CAN-14-1596.

27. Liu, W., Lindberg, J., Sui, G., Luo, J., Egevad, L., Li, T., Xie, C., Wan, M., Kim, S.T., Wang, Z., et al. (2012). Identification of novel CHD1-associated collaborative alterations of genomic structure and functional assessment of CHD1 in prostate cancer. Oncogene *31*, 3939–48. doi:10.1038/onc.2011.554.

28. Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., et al. (2012). Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. *44*, 685–9. doi:10.1038/ng.2279.

29. Faisal, F.A., Sundi, D., Tosoian, J.J., Choeurng, V., Alshalalfa, M., Ross, A.E., Klein, E., Den, R., Dicker, A., Erho, N., et al. (2016).

Racial variations in prostate cancer molecular subtypes and androgen receptor signaling reflect anatomic tumor location. Eur. Urol. *70*, 14–17. doi:10.1016/j.eururo.2015.09.031.

30. Mao, X., Yu, Y., Boyd, L.K., Ren, G., Lin, D., Chaplin, T., Kudahetti, S.C., Stankiewicz, E., Xue, L., Beltran, L., et al. (2010). Distinct genomic alterations in prostate cancers in chinese and western populations suggest alternative pathways of prostate carcinogenesis. Cancer Res. *70*, 5207–12. doi:10.1158/0008-5472.CAN-09-4074.

31. Ren, S., Wei, G.H., Liu, D., Wang, L., Hou, Y., Zhu, S., Peng, L., Zhang, Q., Cheng, Y., Su, H., et al. (2017). Whole-genome and transcriptome sequencing of prostate cancer identify new genetic alterations driving disease progression. Eur. Urol. *73*, 322–339. doi:10.1016/j.eururo.2017.08.027.

32. Boysen, G., Barbieri, C.E., Prandi, D., Blattner, M., Chae, S.S., Dahija, A., Nataraj, S., Huang, D., Marotz, C., Xu, L., et al. (2015). SPOP mutation leads to genomic instability in prostate cancer. Elife *4*, e09207. doi:10.7554/eLife.09207.

33. Cooper, C.S., Eeles, R., Wedge, D.C., Van Loo, P., Gundem, G., Alexandrov, L.B., Kremeyer, B., Butler, A., Lynch, A.G., Camacho, N., et al. (2015). Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. Nat. Genet. *47*, 367–372.

34. Mao, X., Yu, Y., Boyd, L.K., Ren, G., Lin, D., Chaplin, T., Kudahetti, S.C., Stankiewicz, E., Xue, L., Beltran, L., et al. (2010). Distinct genomic alterations in prostate cancers in Chinese and Western populations suggest alternative pathways of prostate carcinogenesis. Cancer Res. *70*, 5207–5212.

35. Li, H., Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*, 589–595.

36. Ye, K., Schulz, M.H., Long, Q., Apweiler, R., Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

37. Hieronymus, H., Schultz, N., Gopalan, A., Carver, B.S., Chang, M.T., Xiao, Y., Heguy, A., Huberman, K., Bernstein, M., Assel, M., et al. (2014). Copy number alteration burden predicts prostate cancer relapse. Proc. Natl. Acad. Sci. U.S.A. *111*, 11139–11144.

38. Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.

39. Farmery, J.H.R., Smith, M.L., NIHR BioResource - Rare Diseases, Lynch, A.G. (2018). Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. Sci. Rep. *8*, 1300.

40. Muthén, B. (2004). Latent variable analysis. The SAGE handbook of quantitative methodology for the social sciences *345*, 106–109.

41. Lee, D.D., Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788.

42. Ghahramani, Z., Griffiths, T.L. (2006). Infinite latent feature models and the Indian buffet process. In: Advances in neural information processing systems. pp. 475–482.

43. Hinton, G.E., Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507.

44. Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. Neural Comput. *14*, 1771–1800.

45. Tran, T., Phung, D., Venkatesh, S. (2011). Mixed-variate restricted Boltzmann machines. In: Asian conference on machine learning. pp. 213–229.

46. Srivastava, N., Salakhutdinov, R.R. (2012). Multimodal learning with Deep Boltzmann Machines. In: Advances in neural information processing systems. pp. 2222–2230.

47. Nguyen, T.D., Tran, T., Phung, D., Venkatesh, S. (2013). Learning parts-based representations with nonnegative restricted Boltzmann machine. In: Asian conference on machine learning. pp. 133–148.

48. Cueto, M.A., Morton, J., Sturmfels, B. (2010). Geometry of the restricted Boltzmann machine. Contemp. Math. *516*, 135–153.

49. LeCun, Y.A., Bottou, L., Orr, G.B., Mülle, K.R. (2012). Efficient backprop. In: Neural networks: Tricks of the Trade. Springer. pp. 9–48.

50. Welling, M., Rosen-Zvi, M., Hinton, G.E. (2005). Exponential family harmoniums with an application to information retrieval. In: Advances in neural information processing systems. pp. 1481–1488.

51. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H. (2007). Greedy layer-wise training of deep networks. In: Advances in neural information processing systems. pp. 153–160.

52. Tran, T., Nguyen, T.D., Phung, D., Venkatesh, S. (2015). Learning vector representation of medical objects via EMR-driven nonnegative restricted boltzmann machines (eNRBM). J. Biomed. Inform. *54*, 96–105.

53. Cui, Z., Ge, S.S., Cao, Z., Yang, J., Ren, H. (2015). Analysis of different sparsity methods in constrained RBM for sparse representation in cognitive robotic perception. J. Intell. Robot. Syst. *80*, 121–132.

54. Wan, L., Zeiler, M., Zhang, S., Le Cun, Y., Fergus, R. (2013). Regularization of neural networks using dropconnect. In: International conference on machine learning. pp. 1058–1066.

55. Srebro, N., Shraibman, A. (2005). Rank, trace-norm and max-norm. In: International Conference on Computational Learning Theory Springer. pp. 545–560.

56. Breiman, L. (1996). Bagging predictors. Mach. Learn. *24*, 123–140.

57. Prechelt, L. (1998). Early stopping - but when? In: Neural Networks: Tricks of the Trade. Springer. pp. 55–69.

58. Hinton, G.E. (2012). A practical guide to training Restricted Boltzmann Machines. In: Neural networks: Tricks of the Trade. Springer. pp. 599–619.

59. Smith, L.N. (2017). Cyclical learning rates for training neural networks. In: Winter Conf. Appl. Comput. Vis. IEEE. pp. 464–472.

60. Olden, J.D., Joy, M.K., Death, R.G. (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Modell. *178*, 389–397.

61. Olden, J.D., Jackson, D.A. (2002). Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Modell. *154*, 135–150.

62. Garson, G.D. (1991). Interpreting neural-network connection weights. AI Expert *6*, 46–51.

63. Larochelle, H., Mandel, M., Pascanu, R., Bengio, Y. (2012). Learning algorithms for the classification Restricted Boltzmann Machine. J. Mach. Learn. Res. *13*, 643–669.

64. Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Trans. Inf. Theory *37*, 145–151.

65. Rousseeuw, P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput. Appl. Math. *20*, 53–65.

66. Pomerantz, M.M., Li, F., Takeda, D.Y., Lenci, R., Chonkar, A., Chabot, M., Cejas, P., Vazquez, F., Cook, J., Shivdasani, R.A., et al. (2015). The androgen receptor cistrome is extensively reprogrammed in human prostate tumorigenesis. Nat. Gen. *47*, 1346–1351. doi:`10.1038/ng.3419`.

67. Gel, B., Díez-Villanueva, A., Serra, E., Buschbeck, M., Peinado, M.A., Malinverni, R. (2015). RegioneR: An R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics *32*, 289–291. doi:`10.1093/bioinformatics/btv562`.

68. Suzuki, R., Shimodaira, H. (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics *22*, 1540–1542.

69. Metzger, E., Willmann, D., McMillan, J., Forne, I., Metzger, P., Gerhardt, S., Petroll, K., von Maessenhausen, A., Urban, S., Schott, A.K., et al. (2016). Assembly of methylated KDM1A and CHD1 drives androgen receptor-dependent transcription and translocation. Nat. Struct. Mol. Biol. *23*, 132–139.

70. Ansari-Pour, N., Zheng, Y., Yoshimatsu, T.F., Sanni, A., Ajani, M., Reynier, J.B., Tapinos, A., Pitt, J.J., Dentro, S., Woodard, A., et al. (2021). Whole-genome analysis of Nigerian patients with breast cancer reveals ethnic-driven somatic evolution and distinct genomic subtypes. Nat. Commun. *12*, 6946.

71. Luce, R.D. (1959). Individual choice behavior: A theoretical analysis. Wiley.

72. Plackett, R.L. (1975). The analysis of permutations. Journal of the Royal Statistical Society: Series C *24*, 193–202.

73. Mollica, C., Tardella, L. (2014). Epitope profiling via mixture modeling of ranked data. Stat. Med. *33*, 3738–3758.

74. Turner, H.L., van Etten, J., Firth, D., Kosmidis, I. (2020). Modelling rankings in R: The PlackettLuce package. Comput. Stat. *35*, 1027–1057. doi:`10.1007/s00180-020-00959-3`.

75. Anderson, G.M. (1976). Error propagation by the Monte Carlo method in geochemical calculations. Geochim. Cosmochim. Acta. *40*, 1533–1538. doi:`https://doi.org/10.1016/0016-7037(76)90092-2`.

76. Kingman, J.F.C. (2005). Poisson processes. In: Encyclopedia of Biostatistics 6. Wiley Online Library.

77. Jones, M.C. (1993). Simple boundary correction for kernel density estimation. Statistics and computing *3*, 135–146.