



UNIVERSIDADE FEDERAL DO PARANÁ

CLARICE DOS SANTOS GROENEVELD

PERFIS DE ATIVIDADE DE REGULON:
INFERÊNCIA E APLICAÇÕES

CURITIBA

2019

CLARICE DOS SANTOS GROENEVELD

PERFIS DE ATIVIDADE DE REGULON: INFERÊNCIA E APLICAÇÕES

Dissertação apresentada ao Curso de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Bioinformática.

Orientador: Prof. Dr. Mauro Antônio Alves Castro

CURITIBA

2019

G874 Groeneveld, Clarice dos Santos
Perfis de atividade de regulon: inferência e aplicações / Clarice dos Santos
Groeneveld. - Curitiba, 2019.
146 p.: il.

Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de
Educação Profissional e Tecnológica, Curso de Pós-Graduação em
Bioinformática, 2019.
Orientador: Mauro Antônio Alves Castro

1. Enzimas - Regulação. 2. Transcrição genética. 3. Câncer – Aspectos
genéticos. 4. Bioinformática. I. Castro, Mauro Antônio Alves.
II. Título. III. Universidade federal do Paraná.

CDD: 572.8



MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO PARANÁ
SETOR DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

Pós-Graduação em Bioinformática WWW.BIOINFO.UFPR.BR
E-mail: bioinfo@ufpr.br Tel: 41 33614906

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em BIOINFORMÁTICA da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **CLARICE DOS SANTOS GROENEVELD** intitulada: **“Perfis de Atividade de Regulon: Inferência e Aplicações”**, após terem inquirido a aluna e realizado a avaliação do trabalho, são de parecer pela sua aprovacao no rito de defesa.

A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

Curitiba, 26 de março de 2019.

Dr. Mauro Antonio Alves Castro
Presidente/Programa de Pós-graduação em Bioinformática – UFPR

Dr. Bonald Cavalcante de Figueiredo
Avaliador Externo/Centro de Genética Molecular do Câncer em Crianças – CEGEMPAC - UFPR

Dr. Fabio Passetti
Avaliador Interno/Instituto Carlos Chagas – ICC-FIOCRUZ-PR/
Programa de Pós-graduação em Bioinformática-UFPR

À minha mãe, que contruiu meu alicerce.

AGRADECIMENTOS

Gostaria de agradecer, primeiramente, ao meu orientador, o Prof. Dr. Mauro A. A. Castro. Ele, juntamente ao nosso colaborador próximo e meu co-orientador não-oficial, Dr. Gordon Robertson, acreditaram em mim todo minuto e me deram suporte para alcançar meus mais ambiciosos objetivos. *Gordon, I can't thank you enough for believing in me.*

Também gostaria de agradecer minha família do laboratório de Bioinformática, os famigerados *coffeeholics* que conseguiram me converter em uma bebedoura de café. Do laboratório de Bioinformática e Biologia de Sistemas, gostaria de agradecer ao Vinícius, meu primeiro contato direto com bioinformática, à Kelin, parceira de discussões e guia de carreira à Sheyla, parceira dos problemas de Linux, à Carol, subidora de ladeiras de Ribeirão e visitante de Nárnia, e ao Jean e à Jéssica, que levarão nosso legado para novos horizontes. Do programa de pós-graduação como um geral, agradeço ao Prof. Roberto Tadeu pelas conversas mirabolantes, à Suzana, pela ajuda com todas as burocracias que aparecem pelo caminho e especialmente à Mariane, companheira de assuntos sérios e triviais.

Reservo também um espaço especial para agradecer às agências financiadoras não só do meu trabalho, como o de todos os meus colegas e da maior parte da ciência do Brasil - CAPES e CNPq. Sem a bolsa da CAPES, não teria conseguido trabalhar em dedicação exclusiva e a qualidade do meu trabalho não seria a mesma. Sem o CNPq, não teríamos as ferramentas do nosso trabalho - computadores e servidores.

Finalmente, às pessoas que não contribuíram para o meu projeto de mestrado, e sim para o meu projeto de pessoa. Agradeço à minha melhor amiga há quinze anos, Jéssica Jerônimo (*You're my person*). Ao meu pai (*Dad, I'm so happy you're back in my life*). À minha irmã, que torce furiosamente por mim. E, por último, porque último tem o melhor sabor, à minha maior apoiadora, minha mãe, Letícia. Sem você, eu não estaria aqui.

“Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.”

Stephen Few

RESUMO

O entendimento do câncer é um dos desafios cruciais da ciência atual. Em busca desta compreensão, governos e institutos em todo o mundo estão gerando uma abundância de dados biológicos. Dentre os tipos de dados, transcriptomas se provaram úteis, mas de interpretação complicada. Redes transcricionais provêm uma estrutura para organizar os dados transcricionais ao redor de elementos regulatórios. Neste trabalho, descrevo um método para inferir estados regulatórios de uma coorte a partir de dados de transcriptoma. Proponho perfis de atividade de regulon (RAPs), implementados no pacote de linguagem R *RTNsurvival*. O *RTNsurvival* já foi utilizado para criar modelos de desfecho de paciente em câncer de mama e fígado, avaliar subtipos de câncer de bexiga, e entender efeitos de interações entre co-reguladores de alvos em variáveis como sobrevida. Finalmente, eu demonstro como atividade de regulon reflete acessibilidade de cromatina em pontos distais, potencialmente enhancer, de alvos positivos e negativos de regulons em câncer de mama. Perfis de atividade de regulon estão se provando uma importante ferramenta estatística para comparação de estados regulatórios de amostras em uma coorte.

Palavras-chave: Análise de rede regulatória. Estado regulatório. Atividade regulatória. Regulon. Sobrevida.

ABSTRACT

Understanding cancer is one of the crucial challenges of contemporary science. In pursuit of this understanding, governments and institutions around the world are generating a wealth of biological data. Among the datatypes, transcriptomes have proven themselves useful, but are many times unwieldy to interpret. Transcriptional networks provide a framework to organize the transcriptional data around regulatory elements. In this work, I describe a method to infer regulatory state of a cohort from transcriptome data. I propose regulon activity profiles (RAPs), implemented in the R language package *RTNsurvival*. *RTNsurvival* has been used to make models of patient outcome in breast and liver cancer, evaluate bladder cancer subtyping results, and understand interactions effects between co-regulators of targets on a variable like survival. I also demonstrate how regulon activity reflects chromatin accessibility at distal, enhancer points of regulon positive and negative targets in breast cancer. Regulon activity profiles are proving to be an important statistical tool in comparing regulatory states of samples within a cohort.

Keywords: Transcriptional network analysis. Regulatory state. Regulatory activity. Regulon. Survival.

SUMÁRIO

Apresentação	12
1 Introdução Geral	13
1.1 Câncer	13
1.2 Fatores de transcrição	14
1.3 Redes regulatórias transcricionais	16
1.4 Estado de uma rede regulatória	18
1.5 Avaliação de subtipos de câncer: um uso para medidas de atividade de redes regulatórias	19
1.6 Justificativa	20
1.7 Objetivo Geral	21
1.8 Objetivos específicos	21
2 <i>RTNsurvival</i>: uma ferramenta para inferência e avaliação de desfechos com atividade de regulon	22
2.1 Introdução	22
2.2 Descrição da ferramenta	22
2.2.1 Cálculo da atividade: Análise de Enriquecimento de Conjuntos de Genes de Duas Caudas (GSEA2)	24
2.3 Manuscrito	25
2.4 Material suplementar: Estudos de Caso	28
2.5 Funções em desenvolvimento: enriquecimento e diferença entre subgrupos de amostras	45
3 Análise integrada de dados de acessibilidade de cromatina (ATAC-seq) e atividade de regulon	51

3.1	Introdução	51
3.2	Métodos	51
3.3	Resultados	52
3.4	Discussão	54
3.5	Conclusão	55
4	Efeito das interações entre <i>dual</i> regulons	56
4.1	Introdução	56
4.2	Materiais e Métodos	56
4.3	Resultados e Discussão	57
4.4	Conclusão	60
5	Conclusão Geral e Perspectivas	62
	Referências	63
	Anexos	69
	Anexo I - The chromatin accessibility landscape of primary human cancers	69
	Anexo II - <i>RTNduals</i> : An R/Bioconductor package for analysis of co-regulation and inference of dual regulons	84
	Anexo III - The consensus molecular classification of muscle-invasive bladder cancer	101

APRESENTAÇÃO

Esta dissertação está estruturada em quatro capítulos. O **capítulo 1** apresenta uma introdução geral do trabalho com revisão bibliográfica, além da justificativa e objetivos. Os capítulos 2 a 4 são compostos de uma introdução ao tópico, específica para aquela análise, além de seções de resultados e discussão.

O **capítulo 2** trata da ferramenta *RTNsurvival* desenvolvida durante o mestrado e apresenta o manuscrito submetido, os dois estudos de caso suplementares ao manuscrito, e adições posteriores à ferramenta. Os capítulos 3 e 4 mostram aplicações da ferramenta, e de outras ferramentas do grupo, em colaboração. O **capítulo 3** faz uma avaliação da correspondência entre atividade de regulon e acessibilidade de cromatina em regiões distais correspondentes a potenciais sítios de ativação dos genes componentes do regulon. No **capítulo 4**, mostramos a integração entre duas ferramentas criadas pelo grupo de pesquisa para auxiliar na avaliação de *dual regulons*, conjuntos de genes potencialmente co-regulados por dois reguladores de interesse.

Finalmente, o **capítulo 5** retoma os capítulos anteriores e apresenta as conclusões e perspectivas do trabalho.

1 INTRODUÇÃO GERAL

1.1 CÂNCER

Câncer é o nome genérico dado a um grande número de doenças neoplásicas. As características que descrevem um tumor podem ser resumidas nas seis marcas do câncer: sustentação de sinal proliferativo, evasão de supressores de crescimento, ativação de invasão e metástase, habilitação da imortalidade replicativa, indução da angiogênese e resistência à morte celular. (Hanahan e Weinberg, 2011)

A Organização Mundial da Saúde estima que 8,8 milhões de pessoas morreram de câncer em todo o mundo em 2015, o que representa um quinto das mortes naquele ano e a segunda causa de morte mais comum. O custo econômico total - que constitui em tratamento, perda de produtividade e morte prematura - foi de aproximadamente US\$ 1.16 trilhão em 2010. (Stewart et al., 2014)

A biologia molecular do câncer é estudada há décadas, mas novas descobertas que evidenciam a grande complexidade de cada neoplasia continuam sendo reveladas. Células, para sobreviver, precisam funcionar como uma máquina bem regulada, em que sinais são enviados e recebidos entre a célula, o tecido e o ambiente em alta velocidade. Quando há desregulações, genéticas ou epigenéticas, as perturbações no sistema podem gerar neoplasias.

Para caracterizar as perturbações que levam ao câncer, cientistas coletam múltiplos tipos de dados de informação genômica (ex.: sequenciamento), transcriptômica (ex.: expressão de mRNA, microRNA e outros RNAs não-codificantes) e epigenômica (ex.: metilação, acessibilidade de cromatina) e proteômica. Anteriormente, esforços eram focados na descrição de pequenos sistemas biológicos. Para mover para uma visão mais holística dos mecanismos da doença, é necessário realizar análises integrativas destes tipos de dados. A biologia de sistemas utiliza-se de grandes quantidades de dados para construir modelos preditivos que objetivam representar a complexidade biológica (Werner et al., 2014).

Há múltiplos bancos de dados armazenando informação molecular de câncer, cuja informação é pública e livremente acessível. O maior deles é o The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network et al., 2013), que provê dados genômicos, transcriptômicos, epigenômicos e proteômicos de 33 tipos de câncer e mais de 20.000 pacientes.

Uma proposta para estudo de sistemas em câncer é o uso de redes regulatórias, com elementos regulatórios como organizadores da informação biológica.

1.2 FATORES DE TRANSCRIÇÃO

Os elementos regulatórios são os orquestradores da expressão gênica no núcleo. Eles agem controlando a expressão de muitos genes ao mesmo tempo utilizando-se de variados mecanismos como: ligação em sítios enhancers ou silenciadores, modulação da abertura e fechamento da cromatina, alterações químicas do DNA como metilação (Spitz e M Furlong, 2012).

A quantidade de interações com múltiplos genes, bem como o efeito cascata de sua expressão, tornam os elementos regulatórios bons candidatos para organização de redes, bem como potenciais alvos para desenvolvimento de fármacos.

Fatores de transcrição, microRNAs (miRNAs) e RNAs longos não-codificadores (lncRNAs) são os três principais tipos elementos regulatórios em estudo atualmente. Apesar de utilizar-se de mecanismos diversos de modulação da expressão, todos são capazes de controlar a expressão de grupos de genes (Farnham, 2009).

Os fatores de transcrição são os elementos regulatórios mais bem caracterizados dos três tipos mencionados. Eles são proteínas que agem modulando a expressão gênica. De acordo com Farnham (2009), há dois tipos de fatores de transcrição: os gerais e os específicos. Os fatores de transcrição gerais modulam toda a expressão gênica se ligando ao promotor, enquanto os específicos têm sítios de ligação com sequências específicas no DNA chamados. Fatores de expressão gerais se ligam ao promotor, que contém domínios conservados, para aumentar o nível de expressão. Porém, o aumento promovido por esse tipo de fator de transcrição é pequeno (Sandelin et al., 2007) quando comparado aos incrementos na expressão mais significativos que vêm de fatores de transcrição ligando-se a sítios denominados enhancers, como ilustrados na Figura 1.

Os fatores de transcrição gerais (ovais verdes) ligam-se à TATA-box e ao iniciador (INR) na fase 1. Para aumento da expressão, o fator de transcrição trapezoidal vermelho liga-se à um sítio próximo ao promotor, estabilizando o recrutamento dos fatores de transcrição gerais. Um aumento mais significativo da transcrição pode ser atingido quando o fator de transcrição hexagonal laranja

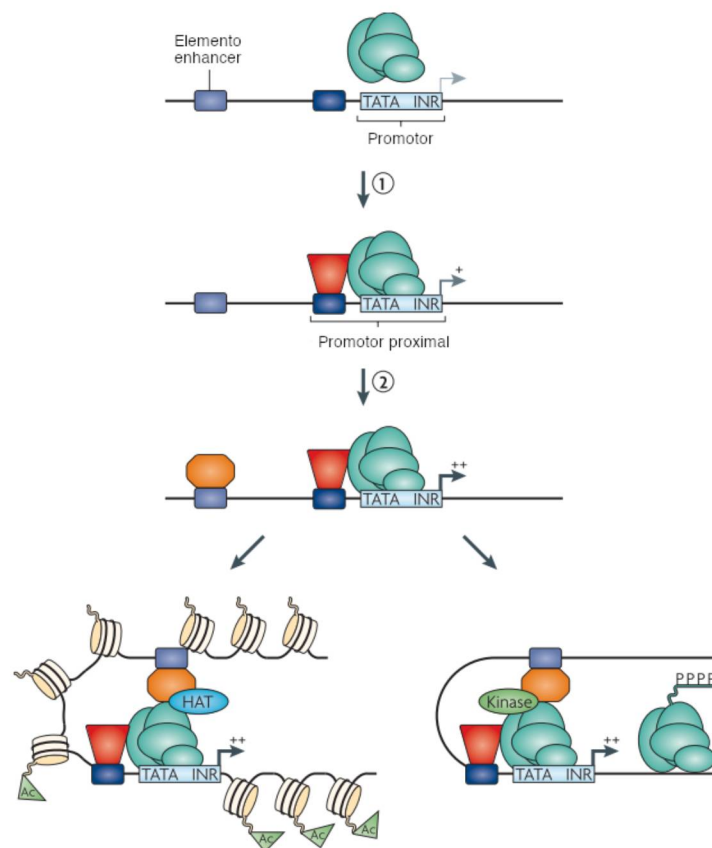


FIGURA 1 – Modelo simplificado de regulação por fatores de transcrição, mostrando o papel de região promotora e enhancers. **Topo:** ligação de fatores de transcrição gerais ao promotor; **centro superior:** ligação de um fator de transcrição específico; **centro inferior:** ligação de fator de transcrição distal; **inferior:** mecanismos de aumento da velocidade de transcrição (1) recrutamento da proteína HAT ou (2) recrutando kinases. FONTE: Farnham (2009), adaptado

liga-se a um sítio enhancer distal, recrutando a proteína HAT (histona acetiltransferase), que promove acetilação da cromatina para sua abertura e favorecimento da transcrição. Outra opção é o recrutamento de uma proteína kinase capaz de fosforilar a RNA polimerase, favorecendo o alongamento da sequência (Farnham, 2009).

Além de sítios enhancers, há sítios de inativação de expressão, também chamados de silencers. A ligação de fatores de transcrição nesses sítios favorecem fechamento da cromatina ou metilação do DNA, diminuindo a expressão do gene (Farnham, 2009).

Estes modos de ação contrários mostram que o aumento da expressão de fatores de transcrição pode estar ligado tanto ao concomitante aumento da expressão de certos genes quanto à diminuição na transcrição de outros. Isto sugere que eles podem ter dois modos de ação, positivo e negativo (Farnham, 2009).

1.3 REDES REGULATÓRIAS TRANSCRICIONAIS

Redes regulatórias transcricionais, também chamadas de redes gênicas regulatórias (GRN, do inglês Gene Regulatory Network), são modelos utilizados para descrever e prever interações biológicas. Segundo Emmert-Streib et al. (2014), há duas famílias de algoritmos para inferência de redes regulatórias transcricionais: (1) os que seguem a perspectiva estatística e (2) os que seguem a perspectiva de modelagem matemática.

Na perspectiva estatística, uma rede é inferida a partir de dados de expressão utilizando algum método de inferência estatística. Nesta rede, nós representam os genes ou produtos de transcrição enquanto as arestas representam interações bioquímicas. Este tipo de rede provê informações gerais sobre interações regulatórias e seus alvos (Matos Simoes et al., 2013).

Já redes derivadas da perspectiva de modelagem matemática têm outro objetivo. Elas em geral são derivadas a partir de conhecimento prévio sobre interações gênicas e têm o objetivo de simular a resposta da transcrição (Emmert-Streib et al., 2014).

Redes derivadas de uma perspectiva estatística são validadas por ChIP-chip ou ChIP-seq – por exemplo, observando-se a ligação de fatores de transcrição específicos na região promotora dos seus alvos preditos – ou sistemas de duplo híbrido (Y2H) para detectar interações proteína-proteína ou

proteína-DNA (Emmert-Streib et al., 2014).

Em sua recente revisão de algoritmos de inferência de redes regulatórias, Delgado e Gómez-Vela (2018) descreveram quatro tipos de métodos para inferência de redes regulatórias transcricionais, baseados em: (1) teoria da informação; (2) redes booleanas; (3) modelos de equações diferenciais (EDOs); (4) modelos bayesianos e (5) modelos neurais.

Dentre estes modelos, os modelos baseados em teoria da informação são os únicos capazes de representar grandes redes de regulação gênica devido a restrições do modelo (e.x. modelos bayesianos) ou restrições de complexidade computacional (redes neurais, ODEs, modelos booleanos).

Os principais algoritmos de inferência de redes regulatórias gênicas baseados em métricas de teoria da informação são o ARACNE (Margolin et al., 2006), CLR (Faith et al., 2007), MRNET (Meyer et al., 2007) e RELEVANCE (Butte e Kohane, 1999).

O algoritmo ARACNE é um dos mais utilizados em biologia computacional do câncer. Ele utiliza informação mútua para inferência de relações regulatórias. Alvos com informação mútua significativa são então filtrados pelo algoritmo DPI (Data Processing Inequality) para identificar as relações regulatórias mais fortes e mais diretas.

Redes regulatórias transcricionais foram utilizadas em alguns trabalhos seminais, incluindo o feito por Carro et al. (2010) na identificação dos iniciadores do fenótipo mesenquimal em gliomas através da inferência de uma rede regulatória tecido-específica e análise de reguladores-mestre da transição entre os fenótipos.

O pacote RTN (Reconstruction of Transcriptional Networks) para linguagem de programação R (Castro et al., 2016) baseia-se na mesmas métricas do ARACNE para computação das redes utilizando-se de uma lista fornecida de potenciais reguladores. Porém, sua implementação fundamentada em permutação e bootstrap tem rigor estatístico e resulta em um ganho em agilidade computacional sobre a implementação original do ARACNE. Redes computadas pelo RTN já foram validadas experimentalmente em câncer de mama (Fletcher et al., 2013).

Recentemente, redes regulatórias organizadas por outros tipos de dados têm ganhado espaço. O aumento na geração de dados de metilação, particularmente pela técnica de sequenciamento de genoma completo marcado com bissulfito (Whole Genome Bisulfite Sequencing, WGBS) possibilitou

a criação de redes regulatórias. A partir dos dados de metilação, é possível prever associações tecido-específicas entre a metilação de regiões do genoma e a expressão de alvos. Juntando essas associações com análises de dados de expressão de reguladores e motivos de ligação de fatores de transcrição, é possível criar um modelo de quais reguladores poderiam estar se ligando nessas regiões enhancer ou silencer preditas (Silva et al., 2018 e Yuan et al., 2018).

1.4 ESTADO DE UMA REDE REGULATÓRIA

A unidade de uma rede regulatória é o regulon: encabeçado por um elemento regulatório e composto por todos os genes cuja expressão tem associação significativa com a expressão do regulador e, portanto, são provavelmente regulados por ele (Castro et al., 2016 e Fletcher et al., 2013).

Regulons são inferidos a partir de dados tecido-específicos com algum método de inferência de redes regulatórias. Eles formam um bloco de anotação funcional específica para um fenótipo (por exemplo, um tipo de câncer), porém estática.

O regulon pode ser tratado como uma anotação tecido-específica composta por listas de genes e reguladores, e pode ser utilizado para entender o estado de uma amostra, conforme é feito com outras anotações funcionais (ex. Gene Ontology, KEGG pathway), a partir de análises de enriquecimento.

O enriquecimento de um regulon, particularmente com um fator de transcrição como regulador, aponta o envolvimento do regulador com o fenótipo (Carro et al., 2010 e Margolin et al., 2006). O enriquecimento do conjunto de genes de um regulon também pode ser tratado como uma medida de sua atividade.

Há diversas medidas utilizadas para inferir o estado regulatório de uma amostra em uma coorte. Vaske et al. (2010) propuseram o algoritmo PARADIGM, que prediz a alteração na atividade de vias canônicas entre dois fenótipos. Já Lefebvre et al. (2010) desenvolveram o MARINa (Algoritmo de Inferência de Reguladores Mestre) que visa reconhecer os fatores de transcrição que controlam a transição entre dois fenótipos (por exemplo, o fenótipo normal e o tumoral) pela atividade do programa completo de transcrição - i.e. de todos os regulons. Mais tarde, o mesmo grupo de pesquisadores propôs uma extensão do MARINa para analisar amostras individualmente, conhecido como ssMARINa (single-sample MARINA) (Aytes et al., 2014).

Duas medidas têm ganhado destaque recentemente para inferência de atividade de regulon. A Análise de Enriquecimento de Conjunto de Genes (do inglês Gene Set Enrichment Analysis, GSEA) foi desenvolvida por Subramanian et al. (2005) para auxiliar a interpretação biológica de dados de expressão de RNA. Ela é frequentemente utilizada para avaliar, por exemplo, o enriquecimento de um termo do Gene Ontology (GO) entre os genes diferencialmente expressos encontrados em um experimento. A GSEA pode ser utilizada para avaliar a atividade de um regulon, especialmente na versão modificada, também conhecida como GSEA de duas caudas (GSEA2), que considera os alvos positivos e negativos de um regulador como dois conjuntos separados (Castro et al., 2016 e Lamb et al., 2006). Outra métrica derivada da GSEA foi implementada por Alvarez et al. (2016) para análise de atividade de regulon dentro de um framework denominado VIPER. O VIPER utiliza uma análise probabilística denominada aREA (Análise de Enriquecimento Baseada em Ranqueamento). Tanto o VIPER quanto a GSEA2 predizem a atividade de cada regulon em cada amostra de uma coorte. Na abordagem GSEA2 proposta por Castro et al. (2016), a média da coorte é utilizada como ponto de referência com o objetivo de distinguir perfis de atividade. Além do trabalho em câncer de mama utilizando a coorte do METABRIC, perfis ternários (atividade positiva, negativa ou indefinida) de atividade regulon foram utilizados para caracterizar subtipos derivados de lncRNAs e miRNAs em câncer de bexiga (Robertson et al., 2017)

1.5 AVALIAÇÃO DE SUBTIPOS DE CÂNCER: UM USO PARA MEDIDAS DE ATIVIDADE DE REDES REGULATÓRIAS

Desde o início dos anos 2000, a identificação de subtipos moleculares em câncer vem sendo utilizada para adicionar uma camada de informação sobre os diferentes tipos de tumores. Sorlie et al. (2003) primeiro identificaram os 5 subtipos moleculares de câncer de mama embasando-se na expressão de genes com baixa variância dentro dos subtipos, mas alta entre eles (Perou e Børresen-Dale, 2011). Subtipagens por assinaturas de genes como a assinatura PAM50 (Prosigna) - que distingue tumores de mama entre os subtipos Luminal A, luminal B, HER2-enriched, basal-like e normal-like - já são utilizadas para guiar decisão clínica, complementando testes patológicos, particularmente para uso em avaliação do prognóstico (Duffy et al., 2017).

O subtipo de cada tumor é mais preditivo de informações como prognóstico e tratamento

do que somente a informação do tecido de origem ou caracterização de imunohistoquímica. Estudos recentes de cortes de pacientes, portanto, têm se trabalhado para determinar subtipos que sejam capazes de exprimir informações sobre a progressão da doença (Perou e Børresen-Dale, 2011 e Robertson et al., 2017).

A determinação de subtipos em câncer se fundamenta no uso de técnicas de clusterização para encontrar agrupamentos de amostras com características moleculares similares. No método de subtipagem PAM50, por exemplo, a expressão de 50 genes medida por técnicas de microarranjo foi utilizada com o algoritmo Particionar ao Redor de Medoids (PAM, uma implementação do algoritmo **k-medoids**). Esse algoritmo agrupa cada amostra junto com amostras a partir do centróide da expressão gênica dos 50 genes marcadores dos subtipos moleculares do PAM50. Com o projeto TCGA e a maior disponibilidade de dados moleculares, técnicas modernas de subtipagem procuram utilizar dados provindos de múltiplas plataformas para formar agrupamentos. Apesar de baseados em diferentes técnicas, os resultados de métodos populares como iCluster (Shen et al., 2013), moCluster (Meng et al., 2016) e Bayesian Consensus Clustering (Lock e Dunson, 2013) tendem a convergir em agrupamentos consistentes (Chauvel et al., 2019).

Subtipagem de tumores respaldada por dados moleculares é um problema de aprendizado de máquina não-supervisionado no qual o objetivo é encontrar classes previamente não conhecidas. Os grupos recuperados pelos métodos de clusterização, então, passam por uma interpretação biológica que procura encontrar semelhanças entre os membros de um grupo e diferenças entre cada grupo e os demais. Por exemplo, os diversos modelos de subtipagem para câncer de bexiga propostos por Robertson et al. (2017) foram avaliados com base em presença de mutações e fusões em genes importantes, características patológicas e histológicas do tumor e também atividade de regulons pré-selecionados validados em uma coorte de pacientes externa.

1.6 JUSTIFICATIVA

Dado a extensa utilidade de métricas de avaliação de estado regulatório como o PARADIGM e o ssMARINa, surge a necessidade de uma forma rápida e automática de realizar o cálculo da atividade de regulon utilizando-se da GSEA2 proposta por Castro et al. (2016).

Aliar esta métrica à análise clássica de sobrevida facilitaria análises de novas coortes e a

proposta de utilização da atividade de regulon como medida pronóstica uma vez que estratificações baseadas em atividade de regulon são melhores do que o perfil de expressão do regulador sozinho. (Castro et al., 2016)

Finalmente, métodos automáticos para avaliação de subtipos de câncer utilizando-se de atividade de regulon, como feito por Robertson et al. (2017), podem auxiliar no entendimento do estado regulatório de subtipos propostos.

1.7 OBJETIVO GERAL

O objetivo geral do trabalho foi desenvolver uma ferramenta para automatizar o uso de atividade de regulon para avaliação de sobrevida e subtipagem em câncer e, em seguida, demonstrar a usabilidade desta ferramenta a partir de estudos de caso.

1.8 OBJETIVOS ESPECÍFICOS

Da ferramenta:

- Desenvolver a ferramenta *RTNsurvival* como um pacote em linguagem de programação R;
- Depositar o pacote no repositório curado de ferramentas biológicas para R, o Bioconductor.

Dos estudos de caso:

- Reconstruir redes regulatórias e avaliar sobrevida com dados regulatórios em diversos tipos de câncer;
- Utilizar a ferramenta *RTNsurvival* para descrever os subtipos;
- Associar atividade de regulon com dados epigenéticos para validar sua utilidade em avaliação de vias regulatórias;
- Propor um modelo centrado em sobrevida com regulons para avaliar motivos de regulação entre dois reguladores;

2 RTNSURVIVAL: UMA FERRAMENTA PARA INFERÊNCIA E AVALIAÇÃO DE DESFECHOS COM ATIVIDADE DE REGULON

2.1 INTRODUÇÃO

A integração de atividade de regulon com modelos de predição de desfechos clínicos, particularmente com análise de sobrevida, fomentou o desenvolvimento da ferramenta *RTNsurvival*. A ferramenta tem estrutura para cálculo de atividade de regulon e para realizar análises integradas de toda a rede regulatória e vários tipos de desfecho. O *RTNsurvival* é a primeira ferramenta capaz de integrar atividade de regulon e análise de sobrevida.

O manuscrito descrevendo a ferramenta e provendo dois estudos de caso (câncer de mama com a coorte METABRIC e câncer de fígado com a coorte TCGA-LIHC) se encontra em revisão na revista *Bioinformatics*.

2.2 DESCRIÇÃO DA FERRAMENTA

A ferramenta foi implementada na linguagem de programação R e está disponível em forma de pacote no repositório de ferramentas para bioinformática do R, o Bioconductor.

O *RTNsurvival* requer a entrada de uma rede transcricional, computada pelo seu pacote parente, o RTN, em forma de um objeto de classe TNI (Transcriptional Network - Inference). Dentro desse objeto, existe a informação de expressão da coorte, anotação dos genes e a inferência da rede regulatória. Opcionalmente, é possível adicionar uma tabela de metadados sobre as amostras ao RTN. Esta tabela é obrigatória para o *RTNsurvival* e, se não estiver disponível no objeto TNI, deverá ser provida separadamente pelo usuário. Dentre os dados da coorte, deve constar um dado de desfecho – por exemplo: morte, morte específica para doença, intervalo livre de progressão, intervalo livre de doença. Este dado precisa estar presente em duas colunas: uma indicando o evento e outra o tempo a partir do tempo 0 da coorte em que o evento ocorreu. O evento deve estar codificado binariamente (1 = o evento ocorreu, 0 = evento não ocorreu). Quando o evento não ocorreu até o tempo marcado, será considerada uma censura daquela nas análises.

O *RTNsurvival* faz pré-processamento dos dados de entrada e cria um objeto de classe TNS

(Transcriptional Network - Survival), que contém todos os dados necessários para realizar a análise.

A segunda etapa do pipeline do *RTNsurvival* é a inferência de perfis de atividade de regulon na coorte. O *RTNsurvival* utiliza duas possíveis métricas para realizar os cálculos de atividade: GSEA2 ou aREA3, com habilidade de utilização de computação paralela para agilizar o cálculo.

Uma vez obtidos os perfis de atividade de regulon, o *RTNsurvival* pode utilizá-lo para realizar análises de desfecho. Duas análises básicas de regulons estão contempladas no pipeline: curvas de Kaplan-Meier e regressão de Cox.

Para traçar uma curva de Kaplan-Meier, primeiro é necessário escolher um regulon para subdividir a coorte em estratos por sua atividade. Por padrão, a coorte é dividida em três estratos, de acordo com o estado do regulon (ativado, reprimido ou indefinido). É possível também dividir em cinco estratos (com estratos representando muito ativado, ativado, indefinido, reprimido e muito reprimido) e sete estratos. Cada estrato irá gerar uma curva no gráfico de Kaplan-Meier. A significância da diferença entre os estratos é avaliada por um teste de log-rank (teste de Mantel-Cox).

Já a regressão de Cox utiliza os valores contínuos de atividade para avaliar a razão de risco da atividade dos regulons. A análise de Cox pode ser feita de forma uni ou multivariada, sendo que covariáveis de interesse podem ser adicionadas pelo usuário (por exemplo: idade, estadiamento, grau do tumor, tamanho do tumor, presença de alguma mutação). A razão de risco é avaliada para um regulon de cada vez, com todas as covariáveis. É possível também avaliar múltiplos regulons ao mesmo tempo.

Quando integrado com uma outra extensão do RTN, o RTNduals, a ferramenta *RTNsurvival* também é capaz de realizar análises de sobrevida em pares de regulons e avaliar o efeito da interação entre regulons em desfechos.

O *RTNsurvival* provê visualizações altamente customizáveis de todas as análises. No pacote, há documentação sobre todas as funções do pipeline bem como uma documentação longa detalhada de todo o processo da ferramenta em forma de vinheta.

2.2.1 Cálculo da atividade: Análise de Enriquecimento de Conjuntos de Genes de Duas Caudas (GSEA2)

O GSEA2 realiza duas análises de enriquecimento de conjunto de genes, considerando os alvos positivos do regulon como um conjunto e os alvos negativos como outro conjunto (*pos* e *neg*, respectivamente). Na computação da rede transcricional no *RTN*, o alvo é indicado como positivo se, além da informação mútua significativa, a expressão do alvo tem coeficiente de correlação (Pearson ou Spearman) positiva com a expressão do regulador, e é indicado como negativo se o oposto é observado. Em cada amostra, é calculada expressão diferencial de cada gene comparando seu valor de expressão na amostra e o valor médio de expressão daquele gene na coorte. Então, os genes são ordenados pelos valores de expressão diferencial. Este rank de genes é denominado o fenótipo. A GSEA2 avalia o enriquecimento de *pos* e *neg* no fenótipo da amostra.

$$P_{hit}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{|r_j|^p}{N_R} \text{ onde } N_R = \sum_{g_j \in S} |r_j|^p \quad (1)$$

$$P_{miss}(S, i) = \sum_{\substack{g_j \in S \\ j \leq i}} \frac{1}{(N - N_H)} \quad (2)$$

Nas equações (1) e (2), queremos avaliar as posições dos genes do o gene set S (alvos positivos ou alvos negativos de um regulon) contendo N_H genes de um universo de N genes. Para todos os genes g_j que pertencem a S , o *Phit* é incrementado de acordo com r_j , que é a distância da expressão do gene g_j na amostra avaliada em relação à referência (média da coorte). Esta correlação é elevada a $p = 1$ e corrigida pelo somatório de todos os valores r pra os genes que pertencem a S . Já para todos os genes que não pertencem a S , *Pmiss* é incrementado.

O escore de enriquecimento (ES, de *enrichment score*) é computado a partir de uma caminhada pelo fenótipo ordenado. Quando um gene do conjunto é encontrado em uma posição do fenótipo, o escore é acrescido (de acordo com o valor P da equação (1)); quando o gene naquela posição do fenótipo não pertence a um conjunto, o escore é decrescido (de acordo com P da equação (2)). O escore de enriquecimento *de facto* é a maior distância entre o eixo x e a curva de enriquecimento, e se dá pela equação (3).

$$ES = P_{hit} - P_{miss} \quad (3)$$

A GSEA funciona em dois casos opostos: ela avalia tanto enriquecimento positivo quanto negativo. Em um conjunto positivamente enriquecido, o escore vai ser positivo uma vez que na caminhada, o algoritmo irá encontrar que os genes do conjunto estão entre os genes positivamente mais diferencialmente expressos na amostra. Um conjunto também pode ser negativamente enriquecido, tendo um escore de enriquecimento negativo. Neste caso, os genes do conjunto estão entre os genes mais negativamente diferencialmente expressos na amostra.

Como dois conjuntos são avaliados (*pos* e *neg*), dois escores de enriquecimento são obtidos: ES_{pos} e ES_{neg} . A atividade do regulon é dada pelo escore diferencial de enriquecimento ($dES = ES_{pos} - ES_{neg}$), cujo valor está entre -2 e 2.

Um valor de dES alto indica um regulon ativado – os alvos positivos estão mais expressos do que na média, enquanto os negativos estão menos expressos do que a média. Isso indica que o regulador está ativamente cumprindo sua função naquela amostra, ativando seus alvos positivos e inativando os negativos. Já um dES baixo indica o oposto – os alvos negativos estão mais expressos do que na média das amostras e os positivos estão menos. O regulador, portanto, não está presente ou não está regulando eficientemente.

2.3 MANUSCRITO

O manuscrito do pacote *RTNsurvival* atualmente se encontra em revisão na revista *Bioinformatics*, como um Application Note. A versão apresentada foi enviada em resposta à primeira rodada de revisão e pode não corresponder completamente ao documento final.

Systems biology

***RTNsurvival*: an R/Bioconductor package for regulatory network survival analysis**

Clarice S. Groeneveld¹, Vinicius S. Chagas¹, Steven J. M. Jones²,
A. Gordon Robertson², Bruce A. J. Ponder³, Kerstin B. Meyer^{3,4} and
Mauro A. A. Castro^{1,*}

¹Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba 81520-260, Brazil, ²Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, BC V5Z4 S6, Canada, ³Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK and ⁴Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, UK Wellcome Sanger Institute, CB10 1SA Hinxton, UK

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on September 9, 2018; revised on February 8, 2019; editorial decision on March 20, 2019; accepted on March 27, 2019

Abstract

Motivation: Transcriptional networks are models that allow the biological state of cells or tumours to be described. Such networks consist of connected regulatory units known as regulons, each comprised of a regulator and its targets. Inferring a transcriptional network can be a helpful initial step in characterizing the different phenotypes within a cohort. While the network itself provides no information on molecular differences between samples, the per-sample state of each regulon, i.e. the regulon activity, can be used for describing subtypes in a cohort. Integrating regulon activities with clinical data and outcomes would extend this characterization of differences between subtypes.

Results: We describe *RTNsurvival*, an R/Bioconductor package that calculates regulon activity profiles using transcriptional networks reconstructed by the *RTN* package, gene expression data, and a two-tailed Gene Set Enrichment Analysis. Given regulon activity profiles across a cohort, *RTNsurvival* can perform Kaplan-Meier analyses and Cox Proportional Hazards regressions, while also considering confounding variables. The [Supplementary Information](#) provides two case studies that use data from breast and liver cancer cohorts and features uni- and multivariate regulon survival analysis.

Availability and implementation: *RTNsurvival* is written in the R language, and is available from the Bioconductor project at <http://bioconductor.org/packages/RTNsurvival/>.

Contact: mauro.castro@ufpr.br

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Transcriptional networks are useful in integrating and interpreting information generated by large-cohort genomics studies. Solutions like *RTN* (reconstruction of transcriptional networks) (Castro *et al.*, 2016) reconstruct these networks, which consist of units made up of a regulatory element and its targets, called regulons. Regulons provide

functional annotations on regulatory associations, and serve as inputs to calculate regulon activity profiles (RAPs) across a cohort. Two recent studies calculated RAPs with the *RTN* package: Castro *et al.* (2016) associated regulon activity with disease-specific survival in breast cancer, and Robertson *et al.* (2017) used regulon status to inform on differences between tumour subtypes in muscle-invasive bladder cancer. While the *RTN* package supports determining regulon

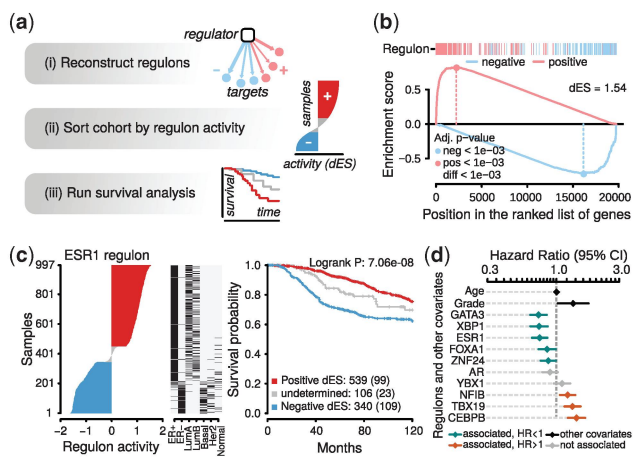


Fig. 1. *RTNsurvival* pipeline and results. (a) Overview of the pipeline. Given a regulatory network from *RTN*, *RTNsurvival* calculates RAPs, which are used to stratify samples for a Kaplan-Meier analysis, or to fit a Cox Proportional Hazards model, including confounding variables. In (b), we show the GSEA-2T regulon activity calculation for sample MB-5365, the luminal A tumour in the METABRIC cohort that has the most-activated ESR1 regulon. The MB-5365 transcriptome is enriched with induced ESR1-positive targets and enriched with repressed ESR1-negative targets. The ‘dES’ score quantifies regulon activity in this sample; GSEA-2T returns one dES per regulon per sample. (c) A covariate and survival analysis for the ESR1 regulon. In the left panel, samples are ranked and stratified according to ESR1 regulon activity. The centre panel adds covariates, and shows that samples with higher ESR1 activity were also found to be ER+ in immunohistochemical assays; such patients were more likely to receive hormone therapy. In the right panel, Kaplan-Meier curves are plotted for the 3 strata. (d) A forest plot generated by a multivariate *RTNsurvival* analysis showing hazard ratios derived from regulon activity for selected regulons, with age and tumour grade

activity, it offers no way to integrate this information with clinical variables. To facilitate such integration, *RTNsurvival* extends *RTN* by combining RAPs with clinical and molecular covariates, and performing uni- or multivariate outcomes analysis, aiding in interpreting differences between subtypes in a cohort.

2 Regulon activity inference and survival analyses

Figure 1a gives an overview of the *RTNsurvival* analysis pipeline. The first step in the pipeline is to infer regulon activity from a transcriptional network. Regulon activity is calculated separately for each sample and regulon, using a two-tailed Gene-Set Enrichment Analysis (GSEA-2T), a modified version of the GSEA-2T approach developed to assess enrichment of two sets of genes (Lamb et al., 2006). The Supplementary Information gives a thorough walk-through of the GSEA-2T metric, and compares it to the related three-tailed analytic Rank-based Enrichment Analysis (aREA-3T) metric (Alvarez et al., 2016). Figure 1b shows the estimation of ESR1 regulon activity for a breast cancer tumour sample from the METABRIC study (Curtis et al., 2012). For each regulon, a cohort can be stratified by activity in order to fit a survival function and generate Kaplan-Meier curves (Kaplan and Meier, 1958). For example, Figure 1c shows three panels for breast cancer tumours from the METABRIC cohort 1. The first panel shows a ranking of cohort’s tumours based on GSEA-2T ESR1 regulon activity, with

the ER-/basal-like samples at the bottom and the ER+/luminal samples at the top. The samples are divided into three groups based on their regulon status (positive dES, undetermined, and negative dES). The second panel shows selected covariates for each tumour, ordered according to the ESR1 regulon activity. The third panel shows a Kaplan-Meier analysis for samples stratified by ESR1 regulon activity. A second survival analysis available in *RTNsurvival* is a Cox Proportional Hazards Model (Cox et al., 1992), which is fit for selected regulons and covariates. Figure 1d shows a forest plot for 10 breast cancer regulons, with covariates age and tumour grade.

3 Case studies

In Section 1 of the Supplementary Information, we apply *RTNsurvival* to explore RAPs and perform survival analysis using the clinical variables from the METABRIC study (Curtis et al., 2012). We calculate RAPs for 997 tumour samples and 36 risk-associated transcription factor regulons, describe the association between regulon activity and subtyping, and report survival results from a Kaplan-Meier analysis and a multivariate Cox regression. In Section 2 of Supplementary Information, for the TCGA hepatocellular cancer (LIHC) cohort (The Cancer Genome Atlas Research Network, 2017), we walk through a similar analysis that uses GRCh38/hg38 harmonized RNA-seq data from the NCI Genomic Data Commons (GDC).

Funding

This work was supported by the National Council for Scientific and Technological Development (CNPq) (407090/2016-9); and the Cancer Research UK (CRUK), the Breast Cancer Research Foundation (BCRF) (BCRF-17-127). C.S.G. and V.S.C. are funded by the Coordination for the Improvement of Higher Education Personnel (CAPES). S.J.M.J. and A.G.R. are funded by the National Cancer Institute of the National Institutes of Health (U24CA210952). The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest: none declared.

References

- Alvarez, M.J. et al. (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat. Genet.*, **48**, 838–847.
- Castro, M.A.A. et al. (2016) Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.*, **48**, 12–21.
- Curtis, C. et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.
- Cox, D.R. et al. (1992) Regression models and life-tables. In: Kotz, S. and Johnson, N.L. (eds) *Breakthroughs in Statistics. Springer Series in Statistics (Perspectives in Statistics)*. Springer, New York, NY.
- Kaplan, E.L. and Meier, P. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.
- Lamb, J. et al. (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
- Robertson, A.G. et al. (2017) Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, **171**, 540–560.
- The Cancer Genome Atlas Research Network. (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, **169**, 1327–1341.

2.4 MATERIAL SUPLEMENTAR: ESTUDOS DE CASO

Por se tratar de um manuscrito no modelo Application Note que é limitado a duas páginas e uma figura, o conteúdo e a discussão do manuscrito se encontram no Material Suplementar. Dois estudos de caso compõe o material suplementar do manuscrito *RTNsurvival*: utilizando a coorte 1 do METABRIC; e a coorte TCGA-LIHC. As duas sessões contém demonstrações das principais funcionalidades do pacote.

Supplementary Information

***RTNsurvival* case studies: regulon activity as a predictor variable in univariate and multivariate survival analyses.**

Clarice S. Groeneveld¹, Vinícius S. Chagas¹, Steven J. M. Jones², A. Gordon Robertson², Bruce A. J. Ponder³, Kerstin B. Meyer^{3,4}, Mauro A. A. Castro^{1,*}.

¹Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba, 81520-260, Brazil.

²Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, V5Z 4S6, Canada.

³Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, United Kingdom. ⁴Wellcome Sanger Institute, Hinxton, CB10 1SA, United Kingdom.

*Corresponding author: mauro.castro@ufpr.br

Contents

1. METABRIC breast cancer cohort 1	2
1.1 Context	2
1.2 Package installation and data sets	2
1.3 Data preprocessing	3
1.4 Regulon activity of individual samples	3
1.5 Regulon activity profiles	3
1.6 Univariate and multivariate survival analyses with <i>RTNsurvival</i>	4
1.7 Identification of proliferation-related regulons	6
1.8 Other metrics for assessing regulator activity	7
2. TCGA hepatocellular carcinoma cohort (TCGA-LIHC)	9
2.1 Context	9
2.2 Download pre-processed data	9
2.3 Inference of the regulatory network with RTN	10
2.4 Univariate and multivariate survival analyses with <i>RTNsurvival</i>	10
3. Conclusions and perspectives	13
Session information	14
Supplementary References	15

1. METABRIC breast cancer cohort 1

1.1 Context

For the METABRIC breast cancer cohort, Castro *et al.* (2016) described a survival analysis that used regulon activity to sort samples in the cohort, which was then stratified and evaluated by Kaplan Meyer (KM) and Cox regression approaches. The authors also described 36 transcription factors (TFs) that were associated with genetic risk of breast cancer. For these 36 TFs, Fletcher *et al.* (2013) reconstructed regulons using the cohort's microarray transcriptome data (Curtis *et al.*, 2012). Our goals in this section are, for the METABRIC cohort 1 (n=997): (1) to estimate regulon activity for these 36 TFs in individual samples, (2) to use regulon activity to sort and stratify the samples, considering sorted covariates, and (3) to assess regulon activity as predictor variable in univariate and multivariate survival analyses.

1.2 Package installation and data sets

The *RTNsurvival* package is available from the R/Bioconductor repository, together with other required packages. Installing and then loading the *Fletcher2013b* data package will make available all data required for this case study.

```
##-- Set the Bioconductor repository
##-- Please make sure to use bioc version >= 3.8 (R >= 3.5)
source("https://bioconductor.org/biocLite.R")
biocVersion()

##-- Install RTNsurvival and other required packages
##-- RTNsurvival (>=1.4.4); Fletcher2013b (>=1.16.0); RTN (>= 2.6.2)
biocLite(c("RTNsurvival", "Fletcher2013b"))
install.packages("pheatmap")

##-- Call packages
library(RTNsurvival)
library(Fletcher2013b)
library(pheatmap)

##-- Load 'rtni1st' data object, which includes regulons and expression profiles
data("rtni1st")
```

The *rtni1st* data also provides clinical and molecular information for 997 samples from the METABRIC cohort 1 (Curtis *et al.*, 2012). The following variables are included in the *rtni1st* data: time to disease-specific death (*time*), event death (*event*), age (*Age*), tumour grade (*Grade*, *G1*, *G2* and *G3*), tumour size (*Size*), lymph nodes (*LN*), ER status from IHC (*ER+* and *ER-*), PAM50 subtypes (*LumA*, *LumB*, *Basal*, *Her2*, and *Normal*), hormone therapy (*HT*) and ethnicity (*Ethnicity*).

```
##-- Check available attributes in 'colAnnotation'
colAnnotation <- rtni.get(rtni1st, what="colAnnotation")
head(colAnnotation)

##-- A list of transcription factors of interest (here, 36 risk-associated TFs)
risk.tfs <- c("AFF3", "AR", "ARNT2", "BRD8", "CBFB", "CEBPB", "E2F2", "E2F3", "ENO1",
             "ESR1", "FOSL1", "FOXA1", "GATA3", "GATAD2A", "LZTFL1", "MTA2", "MYB",
             "MZF1", "NFIB", "PPARD", "RARA", "RB1", "RUNX3", "SNAPC2", "SOX10",
             "SPDEF", "TBX19", "TCEAL1", "TRIM29", "XBP1", "YBX1", "YPEL3", "ZNF24",
             "ZNF434", "ZNF552", "ZNF587")
```

1.3 Data preprocessing

The data preprocessing consists of a single step that creates a `TNS-class` object. This step uses the `tni2tnsPreprocess` function, which requires (1) a transcriptional regulatory network computed by the `RTN` package, and (2) a list of regulators.

```
##-- Create TNS-class object from the 'rtn1st'
tns1st <- tni2tnsPreprocess(tni = rtn1st, regulatoryElements = risk.tfs,
                           time = "time", event = "event", endpoint = 120,
                           keycovar = c("Age","Grade"))
```

1.4 Regulon activity of individual samples

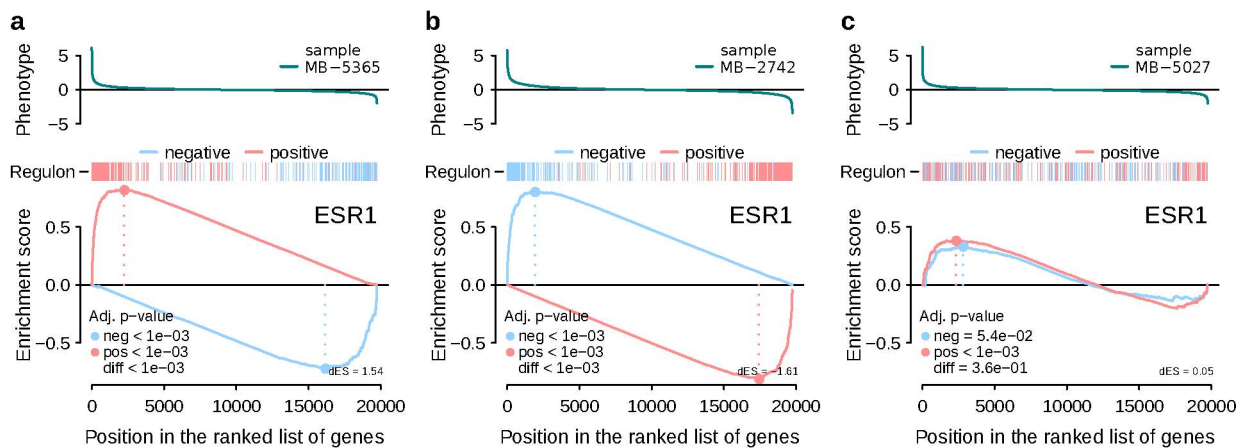
The `tnsPlotGSEA2` function estimates a regulon activity score for a single sample in a cohort, using a two-tailed Gene Set Enrichment Analysis (GSEA-2T). In GSEA-2T, a regulon's positive and negative targets are each considered separate as *pos* and *neg* gene sets. These gene sets are evaluated against a differential gene expression signature, which differs between samples, and is typically calculated in `RTNsurvival` as follows: For each gene in a sample, a differential gene expression is calculated from its expression in the sample relative to its average expression in the cohort; the genes are then ordered as a ranked list representing a differential gene expression signature, also called the sample's phenotype. **Supplementary Figure 1a** shows the estimation of ESR1 regulon activity for a single tumour sample from the METABRIC breast cancer cohort. For each gene set (*pos* and *neg*) a walk down the ranked list is performed, stepwise. When a gene in the gene set is found, its position is marked in the rug plot, with the colour corresponding to the gene set. A running sum, shown as the pink and blue (*pos* and *neg* gene sets, respectively) lines, increases when the gene at that position belongs to the gene set and decreases when it doesn't. The maximum distance of each running sum from the x-axis represents the enrichment score. GSEA-2T produces two per-sample enrichment scores (ES), whose difference ($dES = ES_{pos} - ES_{neg}$) represents the regulon activity. The goal is to assess, for each sample, whether the target genes are overrepresented among the genes that are more positively or negatively differentially expressed. For a sample within a cohort, a large positive dES indicates an *induced* (*activated*) regulon, while a large negative dES indicates a *repressed* regulon. Luminal A sample MB-5365 has an activated pattern for ESR1 (**Supplementary Figure 1a**), while basal-like sample MB-2742 has a repressed pattern (**Supplementary Figure 1b**). The regulon status is assigned as *undetermined* when ES_{pos} and ES_{neg} distributions are skewed to the same side of the ranked list of genes (**Supplementary Figure 1c**).

```
##-- Two-tailed GSEA plots for individual samples
tnsPlotGSEA2(tns1st, "MB-5365", regs = "ESR1")
tnsPlotGSEA2(tns1st, "MB-2742", regs = "ESR1")
tnsPlotGSEA2(tns1st, "MB-5027", regs = "ESR1")
```

1.5 Regulon activity profiles

Regulon activity profiles (RAPs) seek to characterize regulatory program similarities and differences between samples in a cohort. In order to assess a large number of samples, we implemented a function that computes the two-tailed GSEA for the entire cohort. For each regulon, the `tnsGSEA2` function estimates a regulon activity score for each sample in the METABRIC cohort 1.

```
##-- Compute regulon activity for individual samples (this may take a while)
##-- ...for a faster (parallel) option, please see the 'tnsGSEA2' documentation
tns1st <- tnsGSEA2(tns1st)
```



Supplementary Figure 1: Example of using a two-tailed GSEA to calculate ESR1 regulon activity in individual tumour samples. The *phenotype* is the sample's differential gene expression signature, which is obtained by comparing the expression of each gene in the current sample with its average expression across all samples in the cohort. The *phenotype* is used to generate the ranked list of genes on which the two-tailed GSEA is carried out for positive and negative targets (red and blue bars, respectively). For sample PAM50 LumA MB-5365 (a) the ESR1 regulon is activated ($dES > 0$), while for sample PAM50 basal-like MB-2742 (b) the ESR1 regulon is repressed ($dES < 0$). Sample MB-5027 (c) represents an inconclusive case, with positive and negative targets skewed to the same side of the ranked list of genes. These plots reproduce results from Castro *et al.* (2016).

Supplementary Figure 2 shows a heatmap of regulon activity profiles across the METABRIC cohort, together with tumour ER+/- status and PAM50 subtypes. To a large extent, regulon activity segregates samples into meaningful tumour subtypes. These results are consistent with previous studies showing that regulon activity can be used to sort samples in a cohort (for details, examples and additional interpretations on using the dES metric, please refer to Campbell *et al.* (2016), Castro *et al.* (2016), Robertson *et al.* (2017) and Campbell *et al.* (2018)).

```

#-- Get regulon activity and sample attributes
regact_gsea <- tnsGet(tns1st, "regulonActivity")$dif
sdata <- tnsGet(tns1st, "survivalData")
attribs <- c("ER+", "ER-", "LumA", "LumB", "Basal", "Her2", "Normal")

#-- Plot regulon activity profiles
pheatmap(t(regact_gsea), annotation_col = sdata[,attribs], show_colnames = FALSE,
         annotation_legend = FALSE, clustering_method = "ward.D2",
         clustering_distance_rows = "correlation",
         clustering_distance_cols = "correlation")

```

1.6 Univariate and multivariate survival analyses with RTNsurvival

The *RTNsurvival* package uses regulon activity as a predictor variable to study associations between regulons and survival. The `tnsKM` function can be used to generate Kaplan-Meier curves for one covariate (*i.e.* regulon) at a time. **Supplementary Figure 3a** separates the METABRIC cohort ($n=997$ samples) into three strata according to ESR1 regulon activity ($dES < 0$, undetermined, and $dES > 0$), and **Supplementary Figure 3b** shows the corresponding Kaplan-Meier curves. High ESR1 regulon activity is strongly associated with better survival (log-rank $P = 1.96e-08$), reproducing results from Castro *et al.* (2016). **Supplementary Figures 3c-d** illustrate an inverse case, with high PPARD regulon activity associated with poorer survival (log-rank $P = 1.03e-07$). This representation is very convenient for describing the predictor variable along with sample attributes (covariates) and survival curves.


```

#-- Run KM analysis for regulons
tns1st <- tnsKM(tns1st)
tnsPlotKM(tns1st, regs = "ESR1", attribs = attribs, panelWidths=c(3,1,4), width = 6)
tnsPlotKM(tns1st, regs = "PPARD", attribs = attribs, panelWidths=c(3,1,4), width = 6)

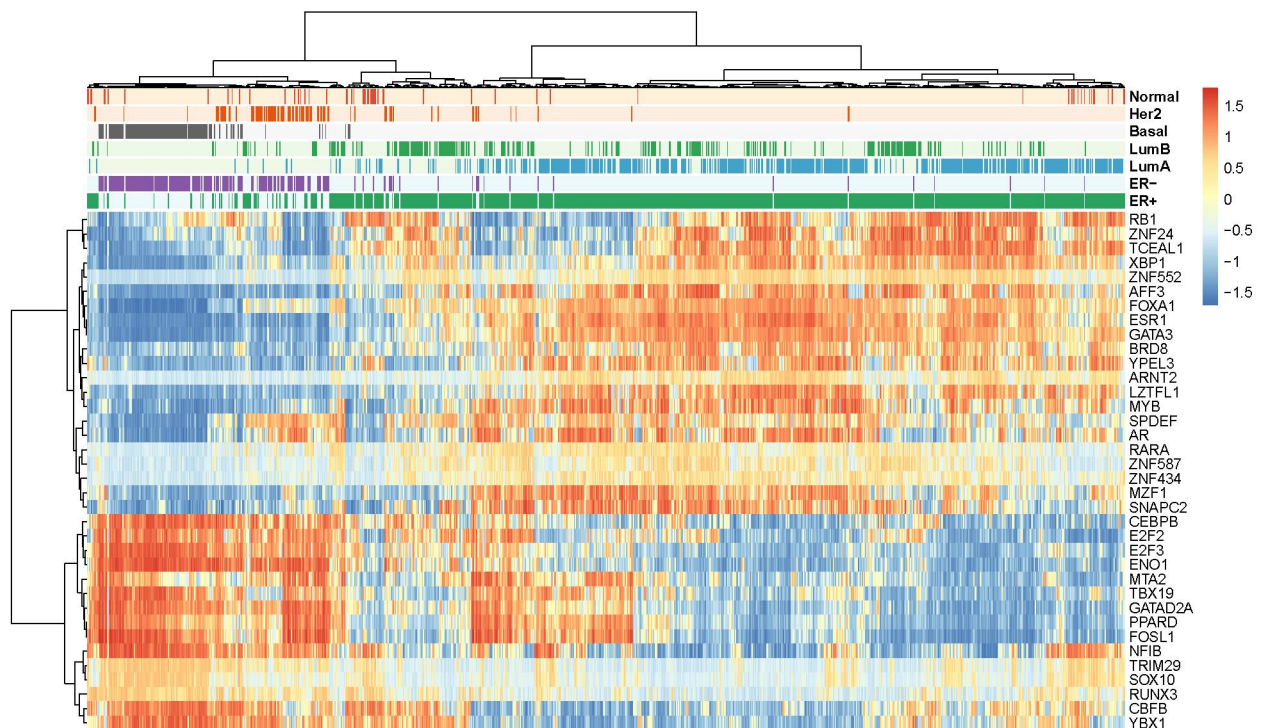
```

Additionally, in order to study the main effects of survival predictors in a multivariate analysis we use the `tnsCox` function, which can adjust the analysis by including confounding factors or other covariates. This function relates the activity of one regulon to times-to-events in a multivariate, additive Cox proportional hazards model, and generates a graphic showing the calculated hazard ratios (HR). **Supplementary Figure 3e** shows that within the 36 regulons there are two subsets with statistically significant hazard ratios (HR < 1 or HR > 1, 95% CI). The regulons associated with higher risk have higher activity values in ER-tumours, particularly basal-like tumors; conversely, regulons associated with lower risk have higher activity in ER+ tumours (**Supplementary Figure 2**).

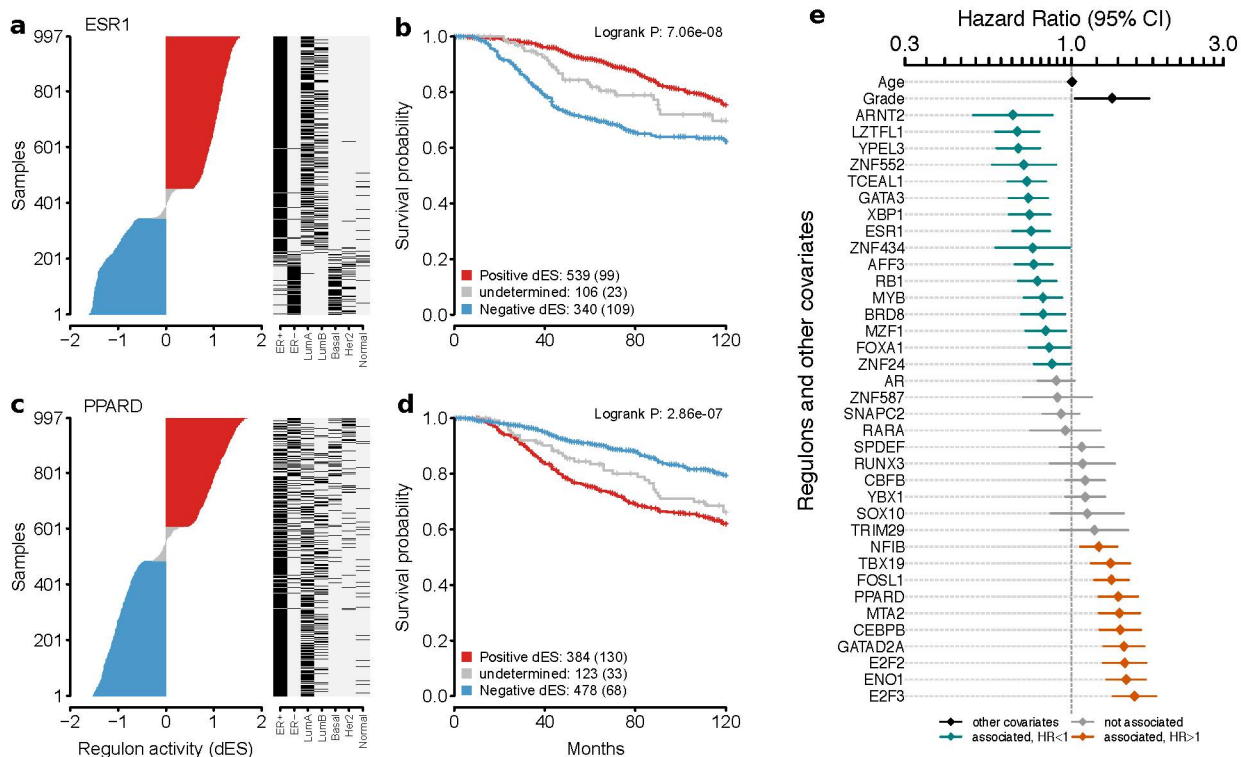
```

#-- Run Cox analysis for regulons
tns1st <- tnsCox(tns1st)
tnsPlotCox(tns1st, height = 7)

```



Supplementary Figure 2: Unsupervised hierarchical clustering of regulon activity profiles across the 997 samples of METABRIC cohort 1 for the set of 36 TFs associated with genetic risk of breast cancer described in Castro *et al.* (2016).



Supplementary Figure 3: Univariate and multivariate survival analyses for single regulons. For the ESR1 regulon: (a) Left: stratification by ESR1 regulon activity (dES) of all 997 samples in METABRIC cohort 1. Samples with inconclusive regulon activity (*i.e.* undetermined status) are indicated in grey. Right: ER status and PAM50 subtypes. (b) Kaplan-Meier survival curves for the dES groups highlighted in (a). Numbers indicate patients in each group and, in curved parentheses, deceased patients (results reproduced from Castro *et al.*, 2016). (c-d) As in (a,b), for the PPARC regulon. (e) Cox multivariate analysis for 36 risk-associated regulons, each considered with age, grade, and regulon activity, for disease-specific survival in the METABRIC cohort 1.

1.7 Identification of proliferation-related regulons

Previous literature has indicated challenges in gene set-based survival analysis. Shimoni (2018) described a “random bias” that was attributed to a large proliferation signature that affects a substantial proportion of the genes in the genome. The author implemented a method that removes the bias by adjusting the gene expression data. The method is largely based on the meta-PCNA signature described by Venet *et al.* (2011), which consists of 131 genes that are associated with proliferation in breast cancer. Shimoni (2018) used the meta-PCNA signature to adjust gene expression for a large number of other cancer types. We used the meta-PCNA signature in our original study (Castro *et al.*, 2016) to identify regulons associated with proliferation in breast cancer, following a method that we described in Fletcher *et al.* (2013). The method consists of an enrichment analysis where we test which regulons are enriched with the meta-PCNA genes. Since the meta-PCNA signature was inferred in breast cancer, we can apply it to the METABRIC cohort.

In this example we show how to identify regulons enriched with the meta-PCNA signature. From our 36 risk TFs, only 3 regulons (E2F2, E2F3 and ENO1) are enriched with the signature. All three are linked to poor outcomes, consistent with their enrichment with proliferation markers. Please refer to Castro *et al.* (2016) and Fletcher *et al.* (2013) for additional details.

```
#-- Load meta-PCNA signature available from Fletcher2013b data package
data("miscellaneous")
```

```

#-- Run MRA analysis pipeline
rtna1st <- tni2tna.preprocess(rtna1st, hits=metaPCNA)
rtna1st <- tna.mra(rtna1st)

#-- Check regulons enriched with meta-PCNA genes
metaPCNA_enriched <- tna.get(rtna1st, what="mra")

```

Table 1: Top 10 regulons enriched with meta-PCNA signature

Regulon	Pvalue	Adjusted.Pvalue
PTTG1	1.0e-49	5.7e-47
FOXM1	1.9e-34	5.5e-32
E2F2	6.1e-24	1.1e-21
E2F8	2.1e-23	2.9e-21
HMGB2	1.7e-16	1.9e-14
ILF2	5.3e-13	5.0e-11
VENTX	5.3e-11	4.3e-09
ZNF395	9.1e-11	6.5e-09
TGIF2	1.1e-10	6.6e-09
PURA	7.0e-10	4.0e-08

```
intersect(metaPCNA_enriched$Regulon, risk.tfs)
```

```
## [1] "E2F2" "E2F3" "EN01"
```

1.8 Other metrics for assessing regulator activity

There are other tools that provide computational infrastructure to explore regulatory networks. Lefebvre *et al.* (2010) and Tarca *et al.* (2009) developed competing methods to infer sample-specific activities of curated pathways, called *PARADIGM* (Pathway Recognition Algorithm using Data Integration on Genomic Models) and *SPIA* (Signaling Pathway Impact Analysis), respectively. Both approaches predict pathway activities in a sample using gene expression and/or other genomic data (*e.g.* copy number alterations). One essential aspect of these approaches is that they have been designed to assess activity of curated pathways, usually represented by sets of genes annotated in a peer-reviewed process dedicated to provide understanding on, *e.g.* cells, organisms and ecosystems. Currently a large number of resources provide reference pathway annotation, for example, *KEGG* (Kanehisa *et al.*, 2016), *Reactome* (Fabregat *et al.*, 2018), *PID* (Schaefer *et al.*, 2009), *Gene Ontology* (The Gene Ontology Consortium, 2017) and *MSigDB* (Liberzon *et al.*, 2015), the latter representing gene set collections that encompass various other curated pathway resources. However, neither of these approaches is designed to reconstruct TF-centric regulons for a tissue of interest, and neither calculates regulon activity on an individual sample basis. To our knowledge, only *RTN* (Castro *et al.*, 2016; Fletcher *et al.*, 2013) and *VIPER* (Alvarez *et al.*, 2016) provide computational infrastructure for that purpose, both tools using the same principles as the *MARINa* algorithm (Lefebvre *et al.*, 2010), which is inspired by the two-tailed GSEA (Lamb *et al.*, 2006). Alvarez *et al.* (2016) compared 12 regulon activity metrics and concluded that the three-tailed analytic Rank-based Enrichment Analysis (aREA-3T) algorithm provides better accuracy and specificity in detecting changes in protein activity after genetic perturbations, closely followed by GSEA-2T. Both GSEA-2T and aREA-3T algorithms are available in *RTNsurvival* for sorting samples in a cohort. **Supplementary Figures 3a,b** show GSEA-2T results for the ESR1 regulon. To calculate similar results using aREA-3T:

```

#-- Compute regulon activity for individual samples using aREA-3T algorithm
tns1st_area <- tnsAREA3(tns1st)

```

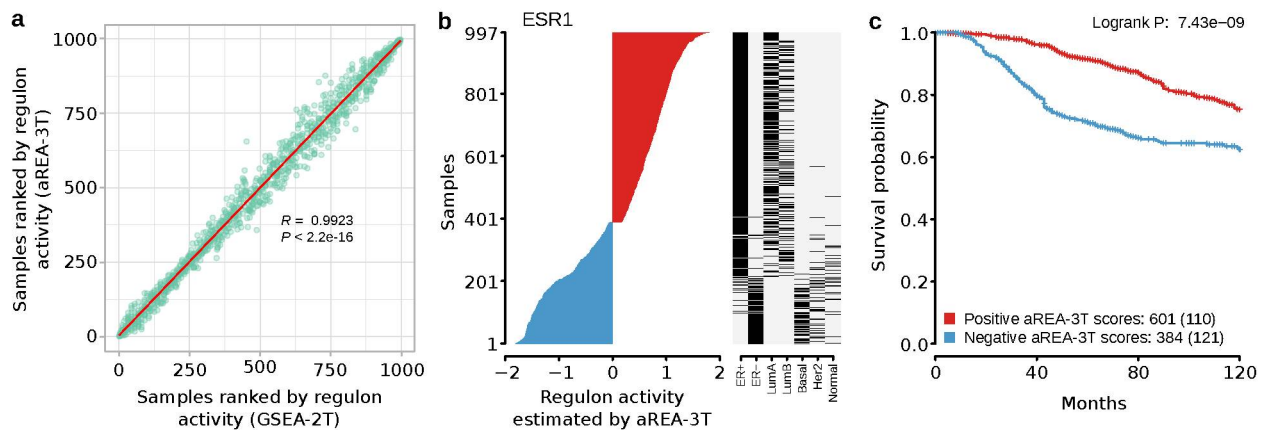
```

#-- Sort sample by regulon activity estimated by aREA-3T and GSEA-2T algorithms
regact_area <- tnsGet(tns1st_area, "regulonActivity")$dif
r_gsea <- apply(regact_gsea, 2, rank)
r_area <- apply(regact_area, 2, rank)
plot(r_gsea[, "ESR1"], r_area[, "ESR1"])

#-- Compute regulon activity for individual samples using aREA-3T algorithm
tns1st_area <- tnsKM(tns1st_area)
tnsPlotKM(tns1st_area, regs = "ESR1", attribs = attribs, panelWidths=c(3,1,4), width = 6)

```

Supplementary Figure 4a shows that aREA-3T and GSEA-2T algorithms are highly concordant in sorting samples by ESR1 regulon activity. **Supplementary Figures 4b,c** show a KM analysis run by RTNsurvival using aREA-3T (compare to Supplementary Figures 3a,b). As the regulon activity scores from the current aREA-3T implementation follow a more continuous distribution than those from GSEA-2T, aREA-3T provides clearer boundaries to stratify the cohort into *pos* vs. *neg* groups, but less-clear boundaries to assign the *undetermined* group; therefore the cohort is simply divided into two groups with positive and negative aREA scores.



Supplementary Figure 4: Concordance between aREA-3T and GSEA-2T algorithms in sorting samples in a cohort. **(a)** The scatter plot shows METABRIC cohort 1 samples ($n=997$) sorted by ESR1 regulon activity estimated by aREA-3T (y-axis) and GSEA-2T algorithms (x-axis). **(b)** Left: stratification by ESR1 regulon activity (estimated by aREA-3T) of all 997 samples in METABRIC cohort 1. Right: ER status and PAM50 subtypes. **(c)** Kaplan-Meier survival curves for the groups highlighted in (b). Numbers indicate patients in each group and, in curved parentheses, deceased patients.

2. TCGA hepatocellular carcinoma cohort (TCGA-LIHC)

2.1 Context

In **section 1**, we used a precalculated transcriptional network for the METABRIC breast cancer cohort, which we made available as the *Fletcher2013b* data package. In **section 2**, we work with a TCGA cohort. We walk through how to use *RTN* and *RTNsurvival* with harmonized GRCh38/hg38 RNA-seq data, which we download from the Genomic Data Commons (GDC, <https://gdc.cancer.gov>) with the *TCGAbiolinks* package (Colaprico *et al.*, 2016). We combine the gene expression data with the cohort's molecular and clinical data, which we download from the The Cancer Genome Atlas Research Network (2017) supplements. We use outcomes data that we download from the Cell web site for the Pan-Cancer Atlas clinical data publication (Liu *et al.*, 2018). We show how to calculate the network from this data with *RTN*, then how to perform outcome analysis with *RTNsurvival*. Our goals are similar to those in **section 1**.

2.2 Download pre-processed data

To run *RTNsurvival* for a new cohort, we need a gene expression matrix for the cohort, a list of transcriptional factors, and patient metadata from the cohort. The patient metadata may consist solely of some outcome — *e.g.* overall survival (OS), progression-free interval (PFI), disease-free interval (DFI). While the patient information must include at least two variables, **time** and **event**, it may also contain more information that can be used as attributes and covariates in *RTNsurvival* functions.

First, we'll download the pre-processed **SummarizedExperiment** object. All the preprocessing steps, from the initial GDC download to the final object, are available on the `csgroen/RTN_example_TCGA_LIHC` repository on Github. The downloaded object consists of three main components: a gene expression matrix, a patient metadata data frame and a gene metadata data frame. We will also get a separate object that contains a list of transcription factors with the necessary annotation.

First, we'll download the pre-processed **SummarizedExperiment** object. All the preprocessing steps, from the initial GDC download to the final object, are available on the `csgroen/RTN_example_TCGA_LIHC` repository on Github. The downloaded object consists of three main components: a gene expression matrix, a patient metadata data frame and a gene metadata data frame. We will also get a separate object that contains a list of transcription factors with the necessary annotation.

```
##-- Repository link and file names
repo_link <- "https://github.com/csgroen/RTN_example_TCGA_LIHC/raw/master/"
fname_exp <- "tcgaLIHCdata_preprocessed.RData"
fname_tfs <- "tfEnsembls.RData"

##-- Download TCGA LIHC data
download.file(paste0(repo_link, fname_exp), fname_exp)
load(fname_exp)

##-- Download transcription factor list and pre-process
download.file(paste0(repo_link, fname_tfs), fname_tfs)
load(fname_tfs)

##-- Call libraries
library(RTNsurvival)
library(SummarizedExperiment)
```

2.3 Inference of the regulatory network with RTN

The *RTN* pipeline starts with the construction of a **TNI-class** object, using the `tni.constructor` method. This method takes in a matrix of gene expression and metadata on the samples and genes, as well as a vector of the regulators to be evaluated. Here, the expression matrix and metadata are available as a `SummarizedExpression` object.

```
##-- TNI constructor
lihcTNI <- tni.constructor(tcgaLIHCdata, regulatoryElements = tfEnsembls)
```

This method also performs pre-processing to check the consistency of all the given arguments and to maximize algorithm performance. It returns a TNI (Transcriptional Network - Inference) object. The next steps run the *RTN* pipeline to generate the regulons (please refer to Fletcher *et al.* (2013), Castro *et al.* (2016) and Robertson *et al.* (2017) for additional details). To run in multithreaded mode, we suggest looking at the `tni.permutation` and `tni.bootstrap` documentation.

```
##-- RTN pipeline
##-- Note: this may take some time; for multithreaded mode, please see
##-- 'tni.permutation' or 'tni.bootstrap' documentation
lihcTNI <- tni.permutation(lihcTNI, pValueCutoff = 10-5, estimator = "spearman")
lihcTNI <- tni.bootstrap(lihcTNI, nBootstraps = 200)
lihcTNI <- tni.dpi.filter(lihcTNI)
```

The `tni.regulon.summary` method lets us get information about the regulons reconstructed by our network. For most calculations, we'll use the DPI-filtered network, which is enriched with direct regulation relationships. From the summary below, we see that the median regulon size is 30 targets and the mean size is about 49, and, while most regulons in the network will be small, some regulons have over 400 targets.

```
tni.regulon.summary(lihcTNI)
```

```
## This regulatory network comprised of 807 regulons.
## -- DPI-filtered network:
## regulatoryElements      Targets      Edges
##           807           17709      39425
##   Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
##     0.0   12.0   30.0   48.9   64.0   434.0
## -- Reference network:
## regulatoryElements      Targets      Edges
##           807           17709     1646659
##   Min. 1st Qu.  Median   Mean 3rd Qu.   Max.
##     0     137   1376   2040   3622   7807
## ---
```

2.4 Univariate and multivariate survival analyses with *RTNsurvival*

For the survival analysis, we'll define Age and Tumour Stage as covariates for the Cox regression and evaluate 5-year (60 months) overall survival (OS).

```
##-- RTNsurvival pipeline
lihcTNS <- tni2tnsPreprocess(lihcTNI,
                             time = "OS.time.months", event = "OS",
                             endpoint = 60, keycovar = c("Age", "Tumor_Stage"))
lihcTNS <- tnsGSEA2(lihcTNS)
```

```
lihcTNS <- tnsKM(lihcTNS)
lihcTNS <- tnsCox(lihcTNS)
```

We can explore the Kaplan-Meier and Cox model results compactly in tables.

```
##-- Explore results
head(tnsGet(lihcTNS, "kmTable"), 10)
```

Table 2: Top 10 regulons in survival curve differences (G-rho test).

Regulons	ChiSquare	Pvalue	Adjusted.Pvalue
FUBP1	35.91304	0.0e+00	0.0000012
TAL1	34.36671	0.0e+00	0.0000013
YBX1	30.82942	0.0e+00	0.0000053
E2F6	29.10570	1.0e-07	0.0000096
HMGA1	32.48896	1.0e-07	0.0000099
ENO1	31.71557	1.0e-07	0.0000107
GMEB1	27.80588	1.0e-07	0.0000107
ETV5	25.72268	4.0e-07	0.0000276
TBX19	23.25694	1.4e-06	0.0000883
TSC22D4	22.56147	2.0e-06	0.0001142

```
head(tnsGet(lihcTNS, "coxTable"), 10)
```

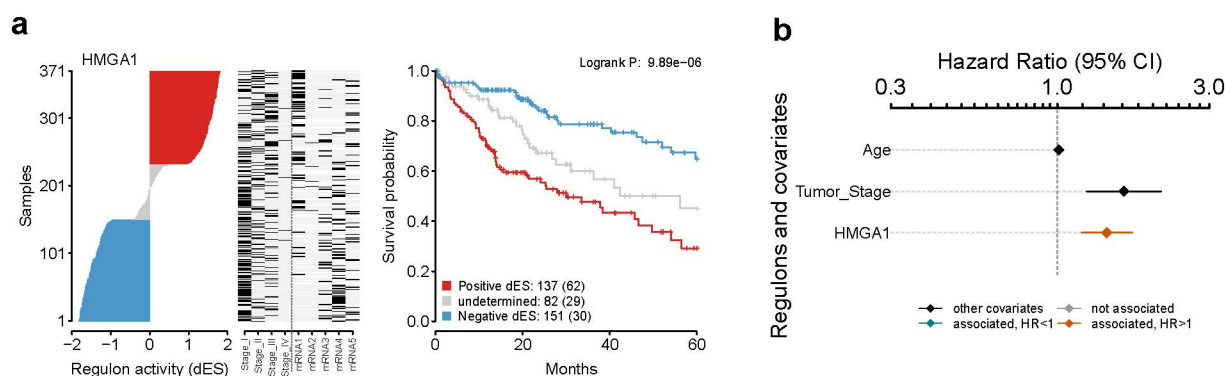
Table 3: Top 10 regulons in Cox Proportional Hazards model.

Regulons	HR	Lower95	Upper95	Pvalue	Adjusted.Pvalue
FUBP1	2.1408242	1.4948719	3.0659003	4.00e-07	0.0002328
YBX1	2.0069439	1.4249945	2.8265538	1.20e-06	0.0003319
HMGA1	1.4307089	1.1916198	1.7177693	2.90e-06	0.0004085
E2F6	2.0915900	1.4349031	3.0488112	2.90e-06	0.0004085
TAL1	0.4488133	0.2940245	0.6850903	6.00e-06	0.0006811
GMEB1	1.9245884	1.3521668	2.7393368	9.40e-06	0.0007730
ZNF408	1.7870618	1.3063348	2.4446946	9.60e-06	0.0007730
Tumor_Stage	1.6178076	1.2370032	2.1158406	1.85e-05	0.0012039
KLF9	0.7333000	0.6161108	0.8727795	2.09e-05	0.0012039
E2F5	1.8708093	1.3156337	2.6602598	2.14e-05	0.0012039

The `tnsPlotKM` method can provide a more complete picture, showing the dynamic range of the activity of a regulon, and how other variables (*e.g.* Stage, mRNA subtypes) are distributed when the cohort is ordered by activity. In this example, we use Tumor Stage and mRNA-cluster membership (only available for the 196 *core* tumour samples, see TCGA, 2017) to get an idea of how samples with low and high HMGA1 activity differ.

```
##-- Kaplan-Meier panel
tnsPlotKM(lihcTNS, "HMGA1",
  attribs = list(c("Stage_I", "Stage_II", "Stage_III", "Stage_IV"),
    c("mRNA1", "mRNA2", "mRNA3", "mRNA4", "mRNA5")),
  panelWidths = c(2,1,3))
##-- Cox multivariate plot
tnsPlotCox(lihcTNS, "HMGA1", ylab = "Regulons and covariates")
```

The left-most panel of **Supplementary Figure 5a** shows the distribution of HMGA1 regulon activity in the cohort tumours, with low activity at the bottom and high activity at the top. The same order is used for



Supplementary Figure 5: Regulon-based survival analysis for HMGA1 in TCGA-LIHC. (a) Three-panel Kaplan-Meier plot for HMGA1. Left: ranking of regulon activity in the samples; Center: Stage and mRNA-cluster covariates along the samples; Right: Kaplan-Meier curve for regulon activity strata. (b) Cox multivariate analysis with covariates Stage, and Age and HMGA1 regulon activity.

the covariate tracks in the center panel, showing tumour stage and mRNA cluster. Given the distribution of the tumours, Stage is an interesting covariate for the Cox model. From **Supplementary Figure 5b**, we see that even when evaluated with Age and Stage, HMGA1 is still informative of survival and linked to increased hazard. In this model, each unit increase in HMGA1's regulon activity corresponds to a 43% higher hazard.

High mobility group A proteins are chromatin remodelers (Sgarra *et al.*, 2018). HMGA1 overexpression induces oncogenesis and metastasis in cultured cell lines of many phenotypes (Sumter *et al.*, 2016). Indeed, its overexpression is also linked to poorer prognostic in several cancer types, including hepatocarcinoma (Chang *et al.*, 2005) (Andreozzi *et al.*, 2016).

For the regulon activity metric, we don't consider the expression of the gene itself, only of its inferred targets; hence, it's a measure of how active a regulator is in a given tumour, not of the regulator's expression in that tumour. Here, we show that in addition to HMGA1's expression being a prognostic marker (see above publications), its regulon activity is also associated with poorer outcomes.

3. Conclusions and perspectives

RTNsurvival extends the functionality of the *RTN* package by finding regulons that are associated with outcomes like survival or progression. The regulon survival analysis uses information about the state of the regulon (*i.e.* the targets of a regulator) to find these associations.

In these examples, we have used transcription factors as examples of regulators. Transcription factors are particularly well-suited for transcriptional networks, but any regulators whose effect can be reliably measured at the transcriptional level can be used by *RTN* and *RTNsurvival*.

While the multivariate analysis provided by the package considers covariates of the user's choice, its default analysis it considers only one regulon at the time with these covariates. (*e.g.* **Supplementary Figure 5b**) For a multivariate survival analysis that considers covariates and more than one regulon at a time, the regulon activity and all relevant covariates can be recovered from the `TNS-class` object, as follows.

```
##-- Get data and bind
full_survData <- tnsGet(lihcTNS, "survivalData")
regulon_activity <- tnsGet(lihcTNS, "regulonActivity")$dif
lihc_data <- cbind(full_survData, regulon_activity)

##-- Example Cox with multiple regulons (FUBP1 and HMGA1)
library(survival)
coxph(Surv(time, event) ~ Tumor_Stage + HMGA1 + FUBP1, data = lihc_data)
```

```
## Call:
## coxph(formula = Surv(time, event) ~ Tumor_Stage + HMGA1 + FUBP1,
##       data = lihc_data)
##
##              coef exp(coef) se(coef)      z      p
## Tumor_Stage  0.5053    1.6574  0.1038  4.869 1.12e-06
## HMGA1        0.1887    1.2077  0.0906  2.083  0.0372
## FUBP1       -0.3651    0.6941  0.1987 -1.837  0.0662
##
## Likelihood ratio test=27.87 on 3 df, p=3.874e-06
## n= 346, number of events= 116
## (25 observations deleted due to missingness)
```

This approach can also be used for more complex survival models, such as LASSO, Adaptive LASSO, Elastic net and others. A LASSO approach was used by Robertson *et al.* (2017) to identify regulons and other covariates linked to outcome in bladder cancer. R packages `hdnom` (Xiao *et al.*, 2016) and `caret` (Kuhn, 2008) provide frameworks for these models.

The current implementation of *RTNsurvival* accepts only regulons identified by *RTN*; for a new cohort we recommend computing regulons with *RTN* (see **section 2**).

Given an *RTN* transcriptional network for a cohort, *RTNsurvival* allows a user to 1) estimate the regulon activity of individual samples, 2) generate regulon activity profiles across a cohort, 3) do univariate and multivariate analyses to associate regulon activity with time-to-event (*i.e.* outcomes) data. Current applications include: 1) assessing covariates across a cohort that has been sorted by regulon activity (Robertson *et al.*, 2017), 2) segregating a cohort for outcomes analysis (Robertson *et al.*, 2017) (Castro *et al.*, 2016), 3) assessing differences between subtypes (Kamoun *et al.*, 2018), and 4) assessing homogeneity/heterogeneity within a subtype (Robertson *et al.*, 2017).

The methods implemented in *RTNsurvival* can also be used with large-scale epigenomic data. For example, recently we showed that regulon activity profiles were consistent with ATAC-seq chromatin accessibility of distal enhancers in breast cancer (Corces *et al.*, 2018). This result provides additional support for regulon activities being a functional readout.

Session information

```
## R version 3.5.2 (2018-12-20)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.1 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/openblas/libblas.so.3
## LAPACK: /usr/lib/x86_64-linux-gnu/libopenblas-r0.2.20.so
##
## attached base packages:
## [1] stats      graphics  grDevices  utils      datasets  methods   base
##
## other attached packages:
## [1] survival_2.43-1      pheatmap_1.0.10      RTNsurvival_1.6.0
## [4] RTNduals_1.6.0       Fletcher2013b_1.18.0 igraph_1.2.2
## [7] RedeR_1.30.0         RTN_2.7.2            Fletcher2013a_1.18.0
## [10] limma_3.38.3
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.0           lattice_0.20-38
## [3] viper_1.16.0         class_7.3-15
## [5] snow_0.4-3           gtools_3.8.1
## [7] digest_0.6.18       GenomeInfoDb_1.18.1
## [9] futile.options_1.0.1 stats4_3.5.2
## [11] evaluate_0.12        e1071_1.7-0
## [13] highr_0.7            gplots_3.0.1
## [15] zlibbioc_1.28.0      VennDiagram_1.6.20
## [17] data.table_1.11.8    gdata_2.18.0
## [19] S4Vectors_0.20.1    Matrix_1.2-15
## [21] rmarkdown_1.11      splines_3.5.2
## [23] BiocParallel_1.16.2 stringr_1.3.1
## [25] mixtools_1.1.0       RCurl_1.95-4.11
## [27] munsell_0.5.0        DelayedArray_0.8.0
## [29] compiler_3.5.2       xfun_0.4
## [31] pkgconfig_2.0.2      BiocGenerics_0.28.0
## [33] segmented_0.5-3.0    htmltools_0.3.6
## [35] SummarizedExperiment_1.12.0 GenomeInfoDbData_1.2.0
## [37] IRanges_2.16.0       matrixStats_0.54.0
## [39] MASS_7.3-51.1        bitops_1.0-6
## [41] grid_3.5.2           gtable_0.2.0
## [43] magrittr_1.5         formatR_1.5
## [45] scales_1.0.0         minet_3.40.0
## [47] KernSmooth_2.23-15   stringi_1.2.4
## [49] XVector_0.22.0       futile.logger_1.4.3
## [51] lambda.r_1.2.3       RColorBrewer_1.1-2
## [53] tools_3.5.2          Biobase_2.42.0
## [55] parallel_3.5.2       yaml_2.2.0
## [57] colorspace_1.3-2     GenomicRanges_1.34.0
## [59] caTools_1.17.1.1     knitr_1.21
```

Supplementary References

- Alvarez,M.J. *et al.* (2016) Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nature Genetics*, **48**, 838–847.
- Andreozzi,M. *et al.* (2016) HMGA1 expression in human hepatocellular carcinoma correlates with poor prognosis and promotes tumor growth and migration in in vitro models. *Neoplasia*, **18**, 724–731.
- Campbell,T.N. *et al.* (2018) ER α Binding by Transcription Factors NFIB and YBX1 Enables FGFR2 Signaling to Modulate Estrogen Responsiveness in Breast Cancer. *Cancer Research*, **78**, 410–421.
- Campbell,T.N. *et al.* (2016) FGFR2 risk SNPs confer breast cancer risk by augmenting oestrogen responsiveness. *Carcinogenesis*, **37**, 741–750.
- Castro,M.A.A. *et al.* (2016) Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature Genetics*, **48**, 12–21.
- Chang,Z. *et al.* (2005) Determination of high mobility group a1 (HMGA1) expression in hepatocellular carcinoma: A potential prognostic marker. *Digestive Diseases and Sciences*, **50**, 1764–1770.
- Colaprico,A. *et al.* (2016) TCGAAbiolinks: An R/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, **44**, e71.
- Corces,M.R. *et al.* (2018) The chromatin accessibility landscape of primary human cancers. *Science*, **362**.
- Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.
- Fabregat,A. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, **46**, D649–D655.
- Fletcher,M.N. *et al.* (2013) Master regulators of FGFR2 signalling and breast cancer risk. *Nature Communications*, **4**, 2464.
- Kamoun,A. *et al.* (2018) The consensus molecular classification of muscle-invasive bladder cancer. *bioRxiv*.
- Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, **44**, D457–D462.
- Kuhn,M. (2008) Building predictive models in R using the caret package. *Journal of Statistical Software, Articles*, **28**, 1–26.
- Lamb,J. *et al.* (2006) The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.
- Lefebvre,C. *et al.* (2010) A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, **6**, 377.
- Liberzon,A. *et al.* (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell System*, **1**, 417–425.
- Liu,J. *et al.* (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, **173**, 400–416.e11.
- Robertson,A.G. *et al.* (2017) Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell*, **171**, 540–556.
- Schaefer,C.F. *et al.* (2009) PID: the Pathway Interaction Database. *Nucleic Acids Research*, **37**, D674–D679.
- Sgarra,R. *et al.* (2018) High mobility group a (HMGA) proteins: Molecular instigators of breast cancer onset and progression. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, **1869**, 216–229.
- Shimoni,Y. (2018) Association between expression of random gene sets and survival is evident in multiple cancer types and may be explained by sub-classification. *PLOS Computational Biology*, **14**, e1006026.
- Sumter,T. *et al.* (2016) The high mobility group a1 (HMGA1) transcriptome in cancer and development.

Current Molecular Medicine, **16**, 353–393.

Tarca,A.L. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

The Cancer Genome Atlas Research Network (2017) Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell*, **169**, 1327–1341.e23.

The Gene Ontology Consortium (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, **45**, D331–D338.

Venet,D. *et al.* (2011) Most random gene expression signatures are significantly associated with breast cancer outcome. *PLoS Computational Biology*, **7**, e1002240.

Xiao,N. *et al.* (2016) Hdnom: Building nomograms for penalized cox models with high-dimensional survival data. *bioRxiv*.

2.5 FUNÇÕES EM DESENVOLVIMENTO: ENRIQUECIMENTO E DIFERENÇA ENTRE SUBGRUPOS DE AMOSTRAS

O primeiro ciclo de desenvolvimento do *RTNsurvival* focou em utilidades na análise de sobrevida, particularmente as duas estatísticas mais populares - curvas de Kaplan-Meier e modelos de regressão de Cox. Com o uso da ferramenta, percebemos a necessidade de adicionar funcionalidades para análise e comparação de subgrupos de amostras, de acordo com subtipos ou características genômicas capazes de dividir a coorte. Embora essas funcionalidades não estejam estritamente ligadas à análise de desfecho, inicialmente proposta para o *RTNsurvival*, elas foram o foco do segundo ciclo de desenvolvimento do pacote devido à sua utilidade abrangente. Duas novas funcionalidades foram incorporadas ao pacote após a submissão do manuscrito: a análise de **enriquecimento** de subgrupos e análise de **diferença** entre subgrupos.

A análise de enriquecimento utiliza o teste exato de Fisher para verificar enriquecimento de alta ou baixa atividade de regulons em subgrupos. Por exemplo, em uma estratificação de 1 seção, a coorte é dividida em ativado (+), reprimido (-) e indefinido para cada regulon, dependendo de sua atividade. A análise de enriquecimento verifica, para cada subgrupo e regulon, se o subgrupo está enriquecido com um estrato (+ ou -) do regulon. Para tanto, uma variável de agrupamento deve estar presente nos metadados das amostras. Exemplos de variáveis de agrupamento são: subtipo, estadiamento, marcadores histológicos e sexo; mas qualquer variável capaz de dividir a coorte em dois ou mais grupos pode ser testada. Uma segunda função utiliza-se dos resultados da análise de enriquecimento para gerar uma visualização em forma de um *heatmap*, como exemplificado na Figura 2.

Existe uma alta correspondência entre o enriquecimento de regulons nos subtipos de câncer de mama entre as duas coortes, com perfis muito similares. Esses resultados remontam os dois grupos principais de regulons encontrados por Castro et al. (2016), de regulons cuja alta atividade está ligada a ER+ (subtipos LumA e LumB) como ESR1, GATA3 e FOXA1, e regulons cuja alta atividade está ligado a ER-. A diferença, no entanto, não se dá apenas entre ER+ e ER-, mas cada subtipo tem seu perfil de atividade distinto e com alta correspondência entre as coortes de descoberta (METABRIC coorte 1) e validação (TCGA-BRCA), mesmo partindo de dados de expressão obtidos por técnicas distintas (microarranjo e RNA-seq, respectivamente).

Essa funcionalidade demonstra como perfis de atividade de regulon podem ser úteis para

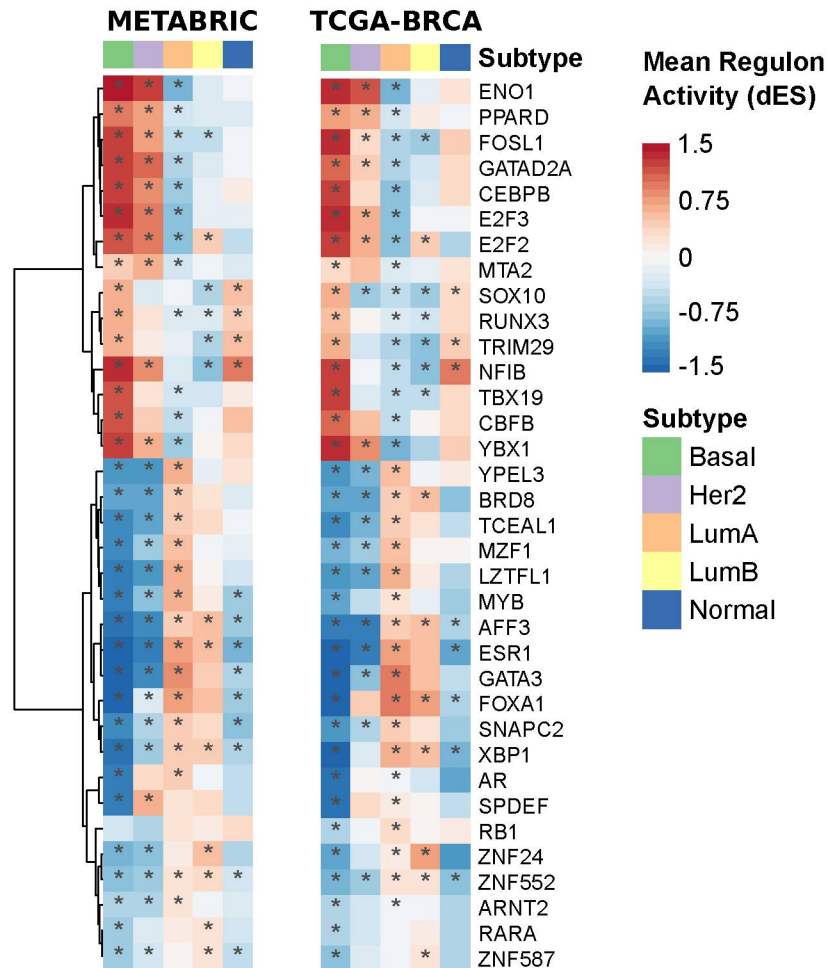


FIGURA 2 – *Heatmap* de média de atividade de 35 TFs de risco por subtipo do PAM50 nas coortes de câncer de mama METABRIC (n=997) e TCGA-BRCA (n=1091), respectivamente. Quadros marcados com asterisco foram significativos no teste de Fisher. Clusterização dos regulons foi feita na coorte METABRIC e replicada na coorte TCGA-BRCA. FONTE: A autora (2019).

avaliar subtipos em diferentes coortes. Os perfis de atividade de regulon foram o foco da colaboração com Kamoun et al. (2018) para caracterização dos subtipos *consenso* de câncer de bexiga.

Além da estatística de enriquecimento, implementamos também uma método para testar diferenças entre subgrupos, que se utiliza de um teste de Wilcoxon (para dois grupos) ou Kruskal-Wallis (para mais de dois grupos) seguido de um pós-teste de Dunn. Todos os tipos de agrupamentos que podem ser testados para enriquecimento também podem ser testados para diferença. Para os subtipos do PAM50 nas coortes METABRIC I e TCGA-BRCA, os resultados dos testes de diferença para 35 dos 36 TFs de risco estão apresentados nas Tabelas 1 e 2.

TABELA 1 – Avaliação da diferença entre subtipos do PAM50 em câncer de mama na coorte METABRIC I. A coluna KW representa o p-valor do resultado do teste de Kruskal-Wallis, enquanto os p-valores específicos para cada subtipo foram calculados pelo pós-teste de Dunn.

Regulon	p-Values					
	KW	LumA	LumB	Basal	Her2	Normal
GATA3	9.68e-116	1.81e-07	1.81e-07	8.70e-03	0.008800	8.80e-03
ESR1	1.35e-115	3.25e-02	3.25e-02	1.69e-01	0.169000	1.73e-02
ENO1	1.61e-108	2.52e-08	9.38e-02	8.43e-02	0.084300	9.38e-02
E2F2	8.74e-108	5.44e-02	2.94e-05	6.36e-02	0.063600	5.44e-02
SOX10	9.96e-107	1.54e-01	1.20e-12	1.63e-01	0.154000	1.63e-01
FOXA1	3.09e-105	3.70e-05	3.70e-05	3.50e-04	0.052700	5.27e-02
TRIM29	2.00e-102	7.65e-04	3.79e-23	7.37e-02	0.000765	7.37e-02
E2F3	7.22e-101	4.61e-05	3.85e-01	9.26e-04	0.000926	3.85e-01
ZNF552	2.02e-100	2.69e-02	2.69e-02	1.90e-02	0.019000	1.75e-02
AFF3	1.67e-99	5.39e-02	5.39e-02	5.60e-02	0.056000	2.65e-03
LZTFL1	2.08e-99	7.83e-13	2.71e-03	6.37e-02	0.063700	2.71e-03
NFIB	7.54e-96	9.88e-12	9.88e-12	6.73e-03	0.101000	1.01e-01
YPEL3	4.70e-94	6.35e-03	2.53e-03	4.86e-01	0.486000	6.35e-03
XBP1	1.27e-93	4.04e-01	4.04e-01	2.06e-07	0.287000	2.87e-01
MYB	9.67e-87	4.63e-10	2.18e-07	3.27e-04	0.258000	2.58e-01

TABELA 1 – Avaliação da diferença entre subtipos do PAM50 em câncer de mama na coorte METABRIC I. A coluna KW representa o p-valor do resultado do teste de Kruskal-Wallis, enquanto os p-valores específicos para cada subtipo foram calculados pelo pós-teste de Dunn. (*continued*)

Regulon	p-Values					
	KW	LumA	LumB	Basal	Her2	Normal
CEBPB	5.32e-85	1.66e-08	2.44e-02	3.27e-04	0.000329	2.44e-02
YBX1	9.22e-84	8.62e-12	1.04e-02	2.05e-06	0.062300	6.23e-02
ARNT2	7.28e-82	2.37e-10	1.02e-02	3.08e-01	0.308000	1.02e-02
GATAD2A	2.43e-81	1.02e-04	3.39e-02	8.57e-02	0.085700	3.39e-02
BRD8	8.99e-76	1.25e-02	1.25e-02	2.25e-01	0.225000	1.36e-05
RARA	1.13e-71	1.39e-06	1.39e-06	1.27e-03	0.017200	1.72e-02
MZF1	1.77e-70	1.40e-06	4.05e-01	6.25e-04	0.000678	4.05e-01
ZNF587	2.21e-68	1.41e-07	1.41e-07	2.67e-02	0.097400	9.74e-02
CBFB	7.67e-68	9.60e-08	1.08e-05	1.75e-05	0.313000	3.13e-01
TCEAL1	1.02e-67	1.53e-01	1.53e-01	9.42e-02	0.094200	1.34e-03
SPDEF	4.13e-67	2.15e-01	2.15e-01	7.16e-07	0.000222	7.16e-07
AR	9.56e-66	9.24e-02	8.15e-03	2.08e-08	0.092400	8.15e-03
SNAPC2	2.20e-62	2.14e-01	2.14e-01	1.19e-02	0.220000	2.20e-01
RUNX3	1.89e-61	8.75e-02	8.75e-02	2.82e-02	0.031300	3.13e-02
FOSL1	2.30e-59	2.16e-01	2.16e-01	1.08e-02	0.010800	1.18e-02
TBX19	1.16e-56	2.23e-01	2.23e-01	8.85e-10	0.222000	2.22e-01
ZNF24	5.02e-50	2.43e-07	2.06e-07	2.44e-01	0.244000	1.68e-01
PPARD	1.55e-49	8.84e-02	3.51e-01	2.23e-01	0.223000	3.51e-01
MTA2	5.93e-31	2.29e-01	4.26e-02	4.40e-02	0.044000	2.29e-01
RB1	2.57e-16	2.44e-01	1.26e-01	1.66e-01	0.166000	2.44e-01

TABELA 2 – Avaliação da diferença entre subtipos do PAM50 em câncer de mama na coorte TCGA-BRCA. A coluna KW representa o p-valor do resultado do teste de Kruskal-Wallis, enquanto os p-valores específicos para cada subtipo foram calculados pelo pós-teste de Dunn.

Regulon	p-Values					
	KW	LumA	LumB	Basal	Her2	Normal
GATA3	3.21e-136	8.77e-09	7.41e-07	1.17e-04	1.94e-01	1.94e-01
ESR1	4.36e-135	3.70e-03	3.70e-03	2.48e-02	2.17e-01	2.17e-01
E2F3	1.12e-125	6.98e-07	2.79e-01	2.00e-06	3.80e-03	2.79e-01
E2F2	7.22e-122	5.11e-02	4.33e-02	3.06e-05	4.33e-02	5.11e-02
FOXA1	2.72e-120	1.21e-02	1.21e-02	5.08e-03	1.30e-04	5.08e-03
ENO1	2.05e-119	5.64e-09	2.09e-02	2.00e-02	2.00e-02	2.09e-02
XBP1	2.05e-119	1.47e-01	1.47e-01	1.40e-03	8.63e-03	8.63e-03
YBX1	8.89e-115	5.50e-04	5.50e-04	6.50e-04	7.92e-02	7.92e-02
TRIM29	2.28e-113	5.67e-05	2.86e-11	1.45e-01	5.67e-05	1.45e-01
SOX10	4.07e-111	3.49e-04	3.00e-02	3.17e-02	3.00e-02	3.17e-02
TCEAL1	4.01e-106	2.52e-04	2.52e-04	4.67e-03	5.80e-02	5.80e-02
GATAD2A	6.06e-106	4.57e-04	4.57e-04	2.46e-06	2.22e-01	2.22e-01
AFF3	1.04e-105	3.13e-01	3.13e-01	3.22e-01	3.22e-01	1.04e-02
FOSL1	1.16e-104	2.53e-01	2.53e-01	2.59e-05	2.83e-01	2.83e-01
NFIB	1.53e-103	2.20e-04	2.20e-04	7.70e-02	3.79e-05	7.70e-02
ZNF552	8.45e-103	4.96e-01	4.96e-01	4.37e-01	3.52e-02	4.37e-01
YPEL3	4.53e-100	4.35e-03	8.41e-02	1.34e-01	1.34e-01	8.41e-02
LZTFL1	1.33e-97	6.48e-09	3.03e-04	9.54e-02	9.54e-02	3.54e-02
CEBPB	6.04e-95	1.41e-07	3.58e-04	4.15e-05	3.68e-01	3.68e-01
TBX19	8.60e-95	2.63e-01	2.63e-01	1.29e-06	3.94e-02	2.11e-03
BRD8	4.96e-87	2.53e-01	2.53e-01	4.94e-01	4.94e-01	2.81e-01
SPDEF	2.54e-86	3.38e-01	3.38e-01	3.02e-06	6.43e-03	2.37e-05

TABELA 2 – Avaliação da diferença entre subtipos do PAM50 em câncer de mama na coorte TCGA-BRCA. A coluna KW representa o p-valor do resultado do teste de Kruskal-Wallis, enquanto os p-valores específicos para cada subtipo foram calculados pelo pós-teste de Dunn. (*continued*)

Regulon	p-Values					
	KW	LumA	LumB	Basal	Her2	Normal
MZF1	1.85e-85	8.50e-04	2.57e-01	4.42e-02	4.42e-02	2.57e-01
AR	6.50e-80	2.73e-01	1.00e-02	6.09e-04	2.73e-01	6.09e-04
CBFB	6.51e-75	1.44e-06	5.42e-02	2.25e-05	2.18e-01	2.18e-01
RUNX3	1.67e-74	5.23e-03	5.23e-03	1.41e-01	3.83e-04	1.41e-01
SNAPC2	1.03e-69	2.42e-03	2.42e-03	1.12e-02	4.17e-01	4.17e-01
ZNF24	1.09e-58	9.49e-05	8.39e-11	9.20e-02	1.01e-04	9.20e-02
PPARD	1.44e-54	4.38e-02	1.39e-01	1.02e-01	1.02e-01	1.39e-01
MYB	4.81e-45	2.73e-05	8.50e-03	3.71e-02	1.67e-01	1.67e-01
ZNF587	3.09e-44	6.65e-02	3.18e-06	4.23e-01	6.65e-02	4.23e-01
RARA	4.10e-44	4.97e-03	4.97e-03	1.19e-01	7.09e-02	1.19e-01
ARNT2	8.86e-39	9.23e-02	9.23e-02	8.56e-02	9.60e-03	8.56e-02
RB1	2.14e-26	6.75e-02	4.80e-01	3.42e-04	1.89e-01	4.80e-01
MTA2	6.28e-22	4.94e-02	4.94e-02	3.36e-01	6.07e-02	3.36e-01

Podemos ver que todos os regulons de fatores de transcrição identificados por Castro et al. (2016) têm poder estatístico de dividir os subtipos, sendo que alguns são mais específicos para um subtipo (ex. GATA3 em Luminal A; E2F2 em Basal-like). As diferenças entre os subtipos encontradas pelo teste de Kruskal-Wallis recapitulam os resultados da análise de enriquecimento. Este teste é adequado para encontrar regulons específicos para um ou um subconjunto de determinados subtipos.

3 ANÁLISE INTEGRADA DE DADOS DE ACESSIBILIDADE DE CROMATINA (ATAC-SEQ) E ATIVIDADE DE REGULON

3.1 INTRODUÇÃO

Um dos principais modos de ação de fatores de transcrição na regulação da expressão gênica é na remodelagem da cromatina. Por exemplo, fatores de transcrição pioneiros como FOXA1 (Mayran e Drouin, 2018) podem ligar-se mesmo à cromatina transientemente aberta, sendo precursores para outros reguladores da expressão. FOXA1 e outros TFs comumente se ligam a enhancers, reconhecendo motivos de ligação.

Há muitos softwares especializados em predição de motivos de ligação de fatores de transcrição, mas nem todos os motivos preditos parecem ter efeito na transcrição. No trabalho de Corces et al. (2018), 404 amostras de diferentes tipos de tumores foram avaliados pela análise de cromatina acessível por transposase utilizando sequenciamento (ATAC-seq). 562.709 elementos de DNA transposase-acessíveis (picos) foram encontrados, incluindo elementos proximais e distais.

Uma das frentes de análise dos dados de ATAC-seq procurou estabelecer ligações entre picos distais e expressão de genes. Foram estabelecidos dois tipos de ligações pico gene: pan-câncer e específico para câncer de mama (BRCA-específico).

Sabendo que existe uma relação próxima entre expressão de genes e abertura da cromatina em região enhancer, uma análise ortogonal utilizando atividade de regulon foi utilizada para reforçar as predições ligações pico-gene e ilustrar a relação entre alvos positivos e negativos e subgrupos de amostras com dES positivo, negativo ou indefinido.

A publicação de Corces et al. (2018) incluiu parte dos resultados desta análise dentre as diversas análises destacadas no manuscrito. O manuscrito está disponível como Anexo I deste documento.

3.2 MÉTODOS

Fletcher e colaboradores (2013) identificaram genes alvo para 539 fatores de transcrição na coorte METABRIC 1. Nós utilizamos esses conjuntos de gene alvo nas n=74 amostras de tumores de

câncer de mama do TCGA para qual também tínhamos dados de ATAC-seq.

Para FOXA1 e ESR1, calculamos perfis de atividade de regulon com o método GSEA de duas caudas, implementado no pacote RTN, nas 74 amostras. Então organizamos as 74 amostras pelo perfil de atividade e geramos um heatmap de acessibilidade de cromatina para as amostras ordenadas utilizando os alvos positivos e negativos do regulon para os quais ligações distais pico-gene tinham sido preditas especificamente para câncer de mama. Picos ligados a múltiplos alvos foram utilizados na avaliação de cada alvo. Quando mais de um pico mapeava para o mesmo alvo, a média de todos os picos era utilizada. Alvos sem ligações distais pico-gene não foram utilizados. Foi utilizada a média do sinal entre replicatas técnicas. Para amostras do TCGA, dados de sobrevida global e covariáveis tumorais foram retiradas de Liu et al. (2018).

Análises de sobrevida foram feitas utilizando o pacote do Bioconductor *RTNsurvival* na coorte completa do TCGA BRCA, para qual nós recalculamos escores de atividade de ESR1 para cada amostra. Para a curva de Kaplan-Meier, nós estratificamos em 3 grupos - dES positivo, indefinido e negativo - e avaliamos diferença entre os grupos para sobrevida global em 5 anos usando o teste log-rank. Adicionalmente, fizemos uma regressão de riscos proporcionais de Cox para solidificar a ligação entre atividade de ESR1 e sobrevida, considerando idade no diagnóstico inicial, estadiamento tumoral e status ER e HER2 como covariáveis. Finalmente, testamos a suposição dos riscos proporcionais do modelo de Cox usando o teste de resíduos de Schoenfeld.

3.3 RESULTADOS

Utilizando a rede regulatória de câncer de mama reconstruída por Fletcher et al. (2013) na coorte de câncer de mama METABRIC I (n=997), nós adequamos à rede aos novos dados de expressão (RNA-seq), provindos de 74 amostras de BRCA para as quais também havia dados de ATAC-seq no TCGA. Os regulons seguem o modelo de um fator de transcrição no centro, alvos positivos e negativos. (Figura 3A)

Nossa expectativa em relação à ligação entre dados de acessibilidade de cromatina e atividade de regulon pode ser simbolizada pela Figura 3B. Uma vez que amostras com dES positivo têm expressão relativamente alta de alvos positivos dentro da coorte, nós antecipamos que a cromatina associada com eles alvos estaria mais acessível, enquanto amostras com dES negativo, que têm

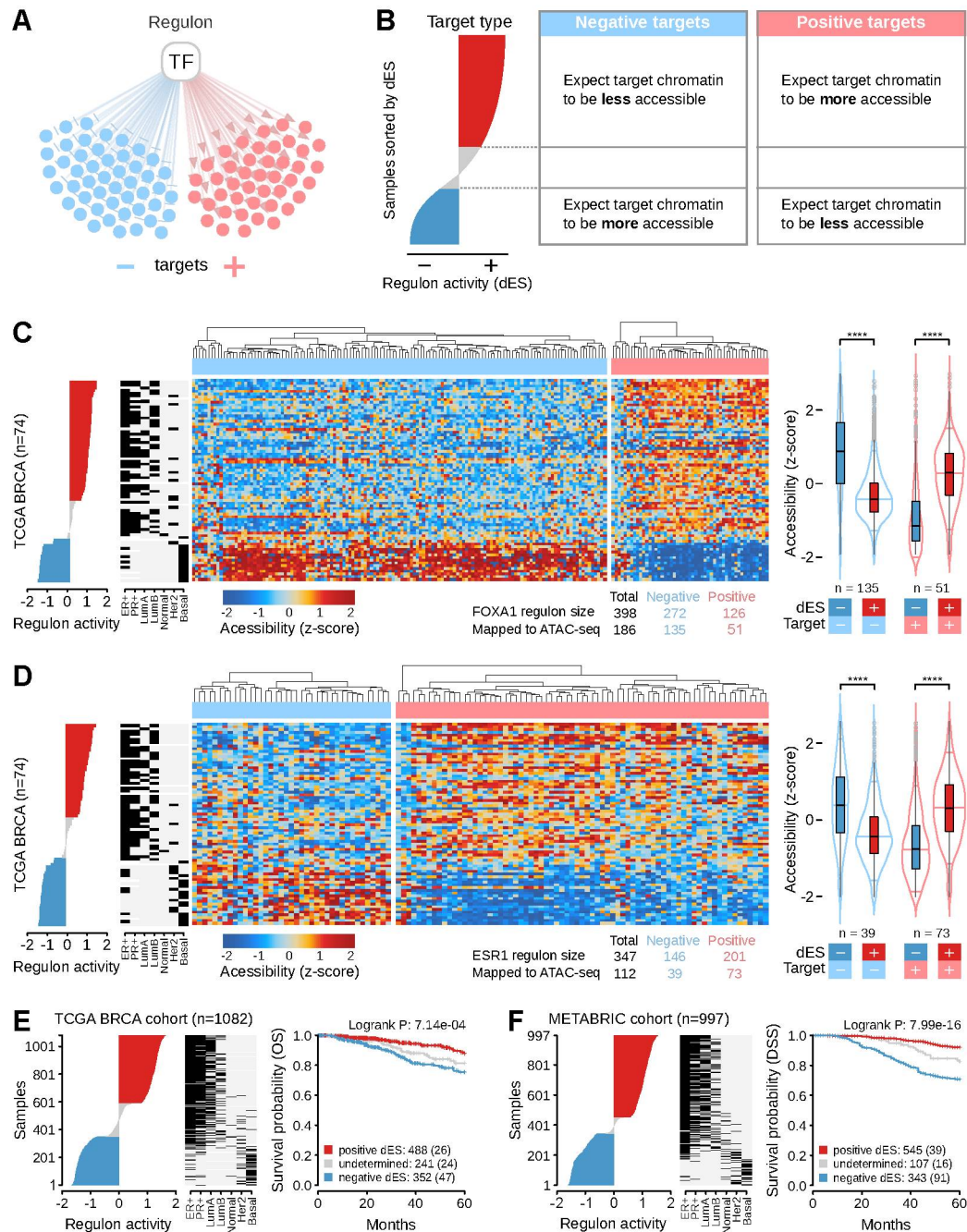


FIGURA 3 – Acessibilidade de cromatina é consistente com expectativas para genes alvos de regulons. **(A)** Representação esquemática de um regulon, mostrando um regulador fator de transcrição e conjuntos de alvos positivos e negativos. **(B)** Esquema das expectativas para acessibilidade de cromatina, para alvos positivos e negativos de um regulador. **(C)** FOXA1 e **(D)** ESR1: Esquerda à direita: perfis de atividade de regulon ordenados (n=74); dados de imunohistoquímica para ER, PR e subtipos moleculares PAM50 ordenados; heatmap de acessibilidade de cromatina ordenado; distribuições de z-score de acessibilidade para alvos negativos/positivos e amostras com dES < 0 e dES > 0. **(E,F)** Ordenação e estratificação da **(E)** coorte completa TCGA BRCA (n = 1082) e **(F)** METABRIC I (n = 997) pela atividade do regulon ESR1. Covariáveis como em (C, D). Para as curvas de Kaplan-Meier, números condizem com amostras tumorais em cada grupo de dES e números entre parênteses são os eventos para sobrevida global e sobrevida específica à doença, respectivamente. FONTE: A autora (2018)

expressão relativamente baixa de alvos positivos, devem ter cromatina ligada a alvos menos acessível. Para alvos negativos, esperamos o oposto.

Testamos essas expectativas para 539 fatores de transcrição (TFs) que tinham pelo menos 15 alvos inferidos na rede. Inicialmente, focamos no TF pioneiro FOXA1 e no receptor de estrogênio ESR1, cuja atividade de regulon está associada à sobrevida em ambas as coortes do METABRIC. (Castro et al., 2016).

Nós calculamos perfis atividade de regulon (dES) para FOXA1 e ESR1 nessas 74 amostras de câncer de mama, bem como na coorte completa do TCGA (n=1082) e coorte 1 completa do METABRIC (n =997). (Figuras 3C a 3F) Nós ordenamos essas amostras e suas covariáveis ER+, PR+ e PAM50 pelo perfil de atividade de cada regulon. Para ambos os regulons, em ambas as coortes, amostras com dES negativo estão enriquecidas no subtipo PAM50 basal e empobrecidas em amostras ER+ e PR+. Para ESR1, grupos de amostras estratificados por estado de regulon negativo, positivo e indefinido tiveram diferenças estatisticamente significativas em sobrevida global para a coorte do TCGA (log-rank $P = 7 \times 10^{-3}$; P da regressão de Cox multivariada = 2×10^{-3} , confirmada por análise de resíduos de Schoenfeld) e sobrevida específica à doença para a METABRIC (log-rank $P = 8 \times 10^{-6}$) (Figuras 3E e 3F).

Heatmaps de acessibilidade para FOXA1 e ESR1 foram consistentes com as expectativas descritas acima e diferenças nas distribuições de acessibilidade para dES > 0 vs. dES < 0 foram estatisticamente significativas ($P < 1 \times 10^{-20}$) (Figuras 3C e 3D).

3.4 DISCUSSÃO

Nós mostramos que redes regulatórias gênicas podem prover informação sobre os efeitos de variação genética herdada (Castro et al., 2016). Também utilizamos redes baseadas em fatores de transcrição (TFs) e seus genes alvos (regulons) para informar diferenças entre subtipos de câncer utilizando múltiplas plataformas analíticas do TCGA (Robertson et al., 2017).

Aqui, descrevemos os cálculos baseados em regulon do RTN e usamos dados de ATAC-seq para esclarecer relações entre dois aspectos centrais de regulons: alvos que estão positivamente ou negativamente correlacionados com o regulador e subgrupos de amostras com atividade de regulon

positiva, negativa ou indefinida. Nas amostras com atividade de regulon positiva, ver acessibilidade mais alta para enhancers distais de alvos positivos é consistente com a indução da expressão de alvos positivos pelo fator de transcrição. Analogamente, em amostras com dES negativo, ver acessibilidade mais alta de alvos negativos é consistente com a inibição indireta pelo TF, por exemplo, suprimindo o recrutamento de outros fatores de transcrição e complexos até o promotor de um alvo negativo – na ausência do TF, a cromatina estaria mais aberta (Yamaguchi et al., 2017). Foxa1 também conhecidamente recruta e interage com TFs co-repressores como Tle3 e Grg3 (Jangal et al., 2014 e Naderi et al., 2012). Quando induzido pelo estrogênio, o produto de ESR1, ERalpha, pode ativar ou reprimir expressão de genes alvo (Heldring et al., 2007).

3.5 CONCLUSÃO

Os resultados da integração entre alvos positivos/negativos de regulon, atividade e acessibilidade de cromatina são condizentes com expectativas teóricas e com o que a literatura propõe sobre o modo de regulação de fatores de transcrição. A análise ortogonal utilizando regulons inferidos em uma coorte de câncer de mama (METABRIC) e dados de expressão de outra coorte com o mesmo fenótipo (TCGA BRCA) mostrou a consistência das anotações encontradas com a adição dos dados de acessibilidade de cromatina.

4 EFEITO DAS INTERAÇÕES ENTRE *DUAL* REGULONS

4.1 INTRODUÇÃO

Além do *RTNsurvival*, há outra ferramenta que estende as funcionalidades básicas do RTN, que recomputa redes regulatórias transcrpcionais. O *RTNduals* (Chagas et al., 2019) (submetido) utiliza a rede transcricional para procurar *dual* regulons - pares de regulons que compartilham alvos e podem conter motivos de co-regulação. Se os reguladores A e B tiveram seus regulons ligados pelo *RTNduals*, o *dual* é denotado por A~B.

O pacote *RTNduals* avalia a reconstrução inicial de regulons, antes da aplicação do filtro Data Processing Inequality do *RTN*, que procura minimizar relações devido a regulação indireta. Nesta rede, chamada de rede de referência, regulons se sobrepõe uns aos outros compartilhando alvos. O *RTNduals* se utiliza do teste exato de Fisher (FET) para testar a sobreposição de regulons e determinar um potencial *dual*. Há principalmente dois tipos de *duals*: concordantes e discordantes. *Duals concordantes* têm o mesmo sentido de regulação para seus alvos; alvos positivos são positivos para ambos, e negativos são negativos para ambos. *Duals discordantes* têm padrões opostos de regulação; para o *dual* A~B, alvos positivos de A são negativos de B, e vice-versa.

Uma forma de avaliar a relevância dos padrões de co-regulação encontrados pelo *RTNduals* pode ser explorada utilizando o *RTNsurvival* e análises de desfecho. Para tanto, implementamos um pipeline alternativo no *RTNsurvival*, que avalia atividade em dual regulons e, particularmente, o efeito da interação entre a atividade dos dois regulons que formam um dual regulon no desfecho.

O manuscrito submetido por Chagas et al. (2019) inclui os resultados desta análise como parte do material suplementar, em um dos estudos de caso. O manuscrito está disponível como Anexo II deste documento.

4.2 MATERIAIS E MÉTODOS

As análises de interação entre dual regulons utilizam o RTN para reconstrução de redes regulatórias, o *RTNduals* para inferência dos duals e o *RTNsurvival* para computação das atividades e análises de interação.

Uma vez que uma rede é reconstruída no RTN, o RTNduals utiliza um único passo para encontrar as associações de dual regulons, criando um objeto da classe MBR (Motifs Between Regulons). O objeto MBR pode ser convertido em TNS (o objeto para análises do *RTNsurvival*) em um passo e, então, o *RTNsurvival* pode proceder com curvas de Kaplan-Meier e regressão de Cox.

Para o Kaplan-Meier de um dual, a coorte é estratificada duas vezes em 3 estratos: ativado (+), indefinido e reprimido (-). As informações dessas duas estratificações são unidas em cinco estratos: ativado para ambos os regulons (+/+), ativado/reprimido (+/-), reprimido/ativado (-/+), reprimido/reprimido (-/-) e indefinido. Uma curva então é computada para cada estrato.

Na regressão de Cox, além de possíveis covariáveis, três termos são avaliados: a atividade do regulon A, a atividade do regulon B e um termo de interação A~B, conforme mostrado na equação (4).

$$h(t) = h_0(t) \exp(\beta_1 A + \beta_2 B + \beta_3 (AB) + \beta_4 x_1 + \dots + \beta_{n+3} x_n) \quad (4)$$

Na qual A e B representam as atividades dos regulons A e B e x_1 a x_n representam as covariáveis. A interação só é significativa para o risco ($h(t)$) se o termo β_3 for significativo.

Para o estudo de caso do manuscrito do *RTNduals*, analisamos possíveis interações dos regulons formados por 36 TFs de risco previamente identificados. Então, escolhemos focar no *dual* regulon com o maior efeito de interação e no *dual* regulon mais significativo identificado pelo pacote *RTNduals* para exemplificar resultados que podem ser obtidos com o uso da ferramenta.

4.3 RESULTADOS E DISCUSSÃO

Para fim de prova de conceito, consideramos dois casos de *dual* regulons calculados utilizando a coorte METABRIC de câncer de mama: ESR1~GATA3 e ESR1~SOX10. Castro et al. (2016) reportou que as atividades de ESR1 e GATA3 diminuem o risco enquanto um aumento de atividade de SOX10 aumenta o risco para sobrevida específica à doença (DSS). No estudo de caso incluído em Chagas et al. (2019), procuramos encontrar *dual* regulons entre os 36 regulons de risco que pudessem interagir no desfecho.

Primeiro, avaliamos o caso do dual concordante ESR1~GATA3, que é o dual de maior significância testado no estudo de caso. Há amplas evidências para o dual ESR1~GATA3. GATA3

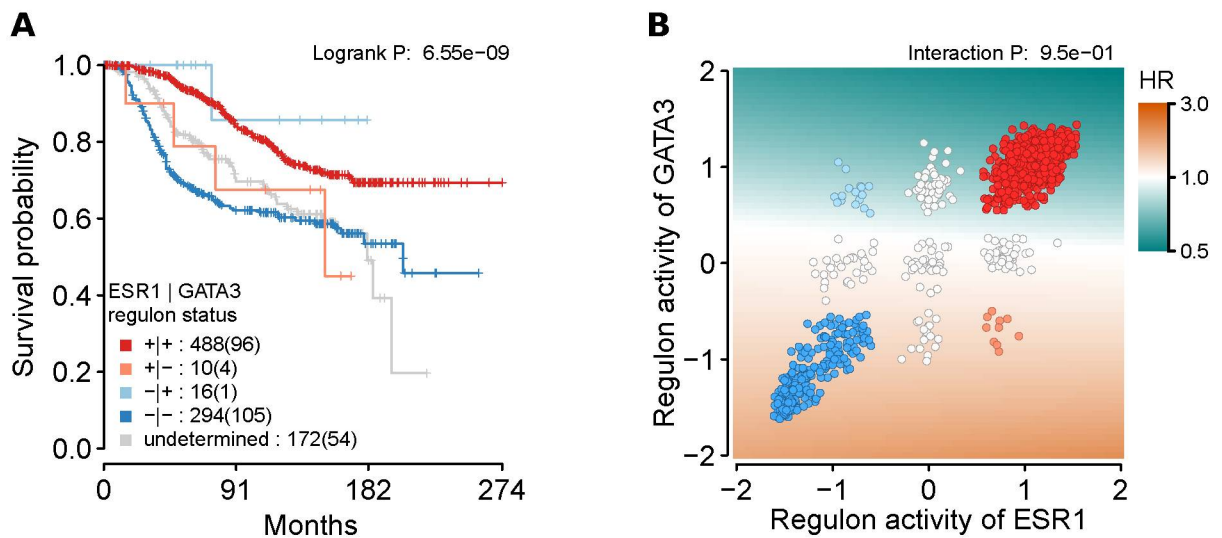


FIGURA 4 – Ausência de interação entre ESR1 e GATA3 em sobrevida. **(A)** Curvas de sobrevida de Kaplan-Meier para o dual regulon ESR1~GATA3. A estratificação de amostras do METABRIC coorte 1 é baseada na combinação do status de atividade dos regulons ESR1 e GATA3. Sinais ‘+’ e ‘-’ indicam $dES > 0$ ou $dES < 0$, respectivamente. Amostras que têm pelo menos um regulon com status inconclusivo são consideradas ‘indeterminadas’. **(B)** Heatmap de interação mostrando o preditor da função de risco variando de acordo com valores de atividade de ESR1 e GATA3. Razões de risco (HR) preditas são representadas por diferentes cores, partindo do verde (baixo risco) até o laranja (alto risco). Pontos representam a distribuição real de valores de dES das amostras presentes na coorte METABRIC 1. As cores dos pontos indicam o estrato destas amostras em (A). Círculos brancos representam amostras indeterminadas em (A). No topo do painel mostramos p-valor ajustado por teste de Benjamini-Hochberg (BH). FONTE: A autora (2019).

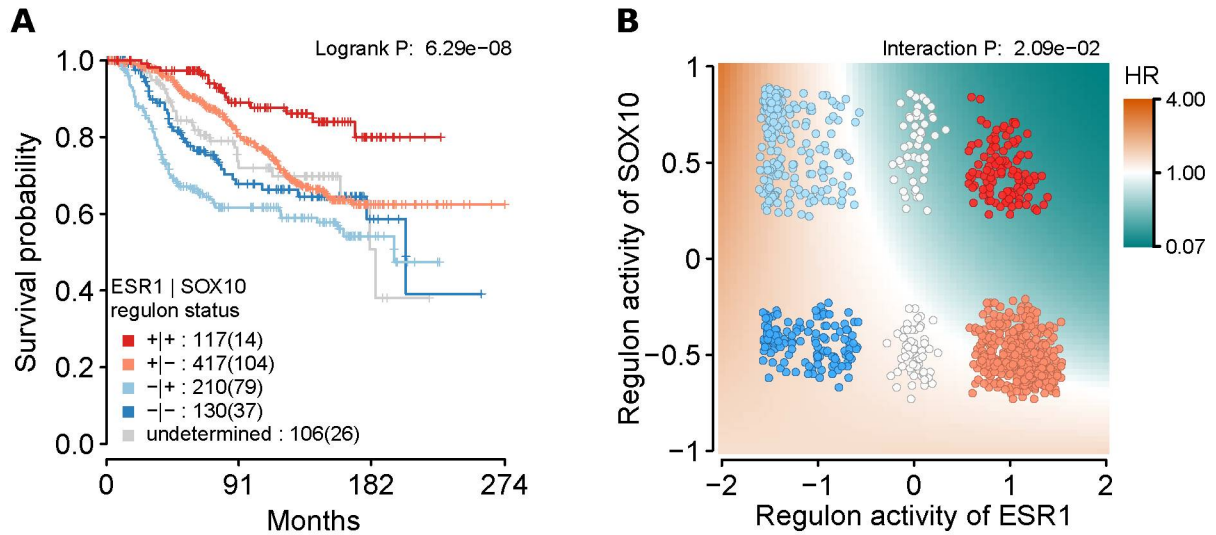


FIGURA 5 – Interação do dual ESR1~SOX10 em sobrevida. **(A)** Curvas de Kaplan-Meier com estratificação baseada no status de atividade de ESR1 e SOX10. **(B)** *Heatmap* de interação no preditor de risco do modelo de Cox. Pontos são as amostras da METABRIC coorte 1 coloridos por seus estratos em (A). FONTE: A autora (2019).

frequentemente coopera com ESR1 mediando acessibilidade em regiões regulatórias. De fato, aproximadamente 40% de todos os eventos de ligação de ESR1 são co-ocupados por GATA3 em linhagem MCF7 (Theodorou et al., 2013).

Apesar de ter alta significância, o dual ESR1~GATA3 não é mais informativo de desfecho do que os regulons ESR1 e GATA3 individualmente. Isto é ilustrado nas Figuras 4 e 6B, que recapitulam o resultado da regressão de Cox – a interação entre ESR1 e GATA3 não é informativa. A estratificação da coorte com base nas atividades dos dois regulons explica a ausência da interação. As atividades de ESR1 e GATA3 são altamente correlatas, e há apenas 26 casos de discordância (perfis +/- ou -/+). Essencialmente, a informação de apenas uma das atividades é suficiente para fazer uma predição sobre risco e a atividade do segundo regulon adiciona muito pouca informação.

Dentre os duals testados, o maior efeito de interação foi encontrado para o dual discordante ESR1~SOX10. Apesar de ser um dual discordante, a estratificação de atividade divide a coorte em cinco estratos com a mesma ordem de grandeza de número de membros. Na curva de Kaplan-Meier mostrada na Figura 5A, vemos que o grupo com o melhor prognóstico tem o perfil de atividade positivo para ESR1 e SOX10 ($++$). Essas amostras têm o perfil consistente com o risco inversamente proporcional à atividade de ESR1 encontrado por Castro et al. (2016). Porém, amostras com a mesma atividade

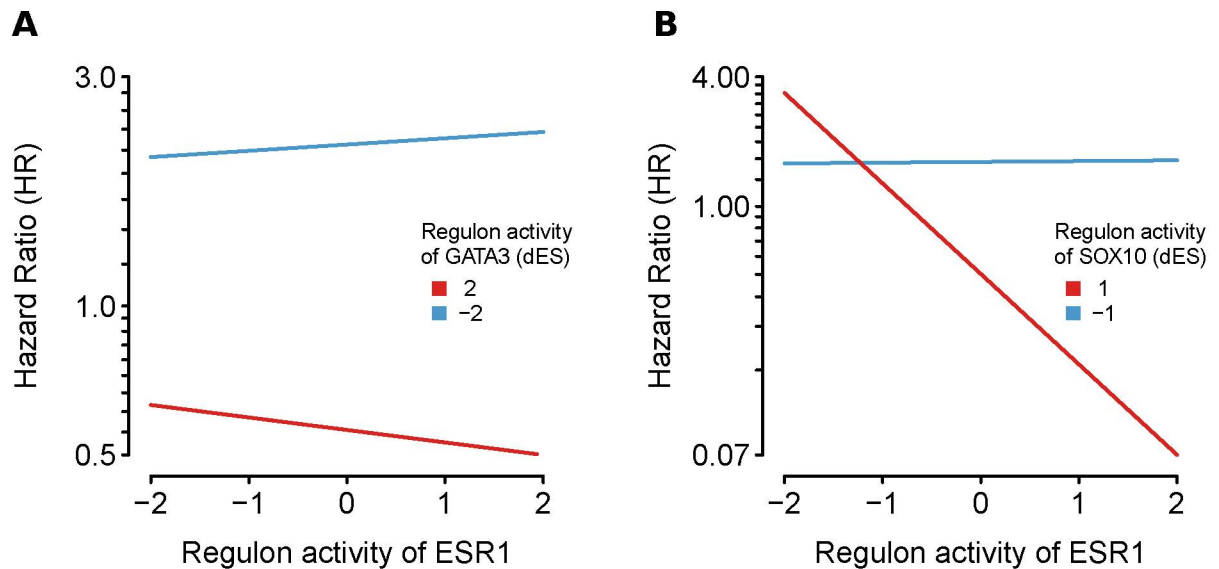


FIGURA 6 – Representação alternativa da interação entre os duals **(A)** ESR1~GATA3 e **(B)** ESR1~SOX10. O eixo x em ambos representa atividade de ESR1. As linhas azuis e vermelhas mostram escores baixos e altos de atividade de (A) GATA3 e (B) SOX10, respectivamente.

para ESR1 e baixa atividade de SOX10 (+/-) têm pior prognóstico do que (+/+), mesmo com a mesma atividade de ESR1. Este perfil mostra o efeito da interação: a adição da informação sobre a atividade de SOX10 permite tirar novas conclusões sobre as amostras classificadas como +/-, os pontos azuis claros da Figura 5B. A Figura 6 recapitula o gráfico de interação mostrado em XB e YB com uma visualização alternativa da interação, mantendo uma variável estável e variando a atividade da outra em dois pontos distintos de atividade da primeira variável. O cruzamento das linhas mostra a existência de uma interação entre ESR1 e SOX10.

4.4 CONCLUSÃO

A integração do *RTNsurvival* com o pacote *RTNduals* provê uma nova linha de análise para refinar a identificação de *dual* regulons, focando em interações entre reguladores que afetam sobrevida. Na análise ilustrativa da METABRIC coorte 1, demonstramos que um *dual* altamente significativo e com ampla evidência externa (ESR1~GATA3) pode não ter efeito de interação, e mostramos como o *dual* ESR1~SOX10 interage na informação de sobrevida. Esta análise está demonstrada no material suplementar de Chagas et al. (2019), como um dos estudos de caso; a primeira versão submetida está

disponível como Anexo II deste documento.

5 CONCLUSÃO GERAL E PERSPECTIVAS

Perfis de atividade de regulon fornecem informações interessantes entre diferenças de amostras na mesma coorte e padrões de regulação em amostras. A integração de métodos de inferência de atividade de regulon com outros tipos de dados enriquece as possibilidades: desde o entendimento da biologia dos subtipos até, potencialmente, identificação de padrões regulatórios complexos no tratamento de tumores com drogas.

Os resultados mostram que a ferramenta *RTNsurvival* desenvolvida durante o projeto pode ser utilizada em vários contextos na análise de coortes de pacientes e sua integração com a ferramenta *RTNduals* traz uma maneira inovadora de avaliar interações entre pares de co-regulação em análise de desfecho.

Regulons e sua atividade foram colocados em teste utilizando duas coortes diferentes para gerar e testar a rede. A concordância entre a estratificação das amostras, bem como na análise de sobrevida de um dos principais regulons do câncer de mama, o *ESR1*, demonstra consistência biológica das relações encontradas. A conformidade entre expectativas e resultados para acessibilidade de cromatina também demonstra que os escores de atividade de regulon representam de forma resumida um padrão biológico que vai além dos dados de expressão que foram utilizados para sua inferência.

As ferramentas *RTNsurvival* e *RTNduals* já foram utilizadas em trabalhos com objetivos de caracterização de subtipos de câncer de bexiga (Kamoun et al., 2018 e Robertson et al., 2017). (Kamoun et al. (2018) está disponível como o Anexo III deste trabalho). Porém, nossas perspectivas são para utilizá-las, e particularmente atividade de regulon, em trabalhos de descoberta visando esclarecer padrões regulatórios e co-regulatórios em câncer.

REFERÊNCIAS

ALVAREZ, M. J.; SHEN, Y.; GIORGI, F. M.; LACHMANN, A.; DING, B. B.; HILDA YE, B.; CALIFANO, A. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. **Nature Genetics**, v. 48, n. 8, p. 838–847, 2016.

AYTES, A.; MITROFANOVA, A.; LEFEBVRE, C.; ALVAREZ, M. J.; CASTILLO-MARTIN, M.; ZHENG, T.; EASTHAM, J. A.; GOPALAN, A.; PIENTA, K. J.; SHEN, M. M.; et al. Cross-Species Regulatory Network Analysis Identifies a Synergistic Interaction between FOXM1 and CENPF that Drives Prostate Cancer Malignancy. **Cancer Cell**, v. 25, n. 5, p. 638–651, 2014. Cell Press.

BUTTE, A. J.; KOHANE, I. S. Mutual Information RELEVANCE Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements. In: *Biocomputing 2000. Anais....* p.418–429, 1999. WORLD SCIENTIFIC.

CANCER GENOME ATLAS RESEARCH NETWORK, J. N.; WEINSTEIN, J. N.; COLLISSON, E. A.; MILLS, G. B.; SHAW, K. R. M.; OZENBERGER, B. A.; ELLROTT, K.; SHMULEVICH, I.; SANDER, C.; STUART, J. M. The Cancer Genome Atlas Pan-Cancer analysis project. **Nature genetics**, v. 45, n. 10, p. 1113–20, 2013. NIH Public Access.

CARRO, M. S.; LIM, W. K.; ALVAREZ, M. J.; BOLLO, R. J.; ZHAO, X.; SNYDER, E. Y.; SULMAN, E. P.; ANNE, S. L.; DOETSCH, F.; COLMAN, H.; et al. The transcriptional network for mesenchymal transformation of brain tumours. **Nature**, v. 463, n. 7279, p. 318–25, 2010. NIH Public Access.

CASTRO, M. A. A.; SANTIAGO, I. DE; CAMPBELL, T. M.; VAUGHN, C.; HICKEY, T. E.; ROSS, E.; TILLEY, W. D.; MARKOWETZ, F.; PONDER, B. A. J.; MEYER, K. B. Regulators of genetic risk of breast cancer identified by integrative network analysis. **Nature Genetics**, v. 48, n. 1, 2016.

CHAGAS, V. S.; GROENEVELD, C. S.; OLIVEIRA, K. G.; TREFFLICH, S.; ALMEIDA, R. C. DE; PONDER, B. A. J.; MEYER, K. B.; JONES, S. J. M.; ROBERTSON, A. G.; CASTRO, M. A. A. RTNduals: An R/Bioconductor package for analysis of co-regulation and inference of dual

regulons. **Bioinformatics - Submitted**, 2019.

CHAUVEL, C.; NOVOLOACA, A.; VEYRE, P.; REYNIER, F.; BECKER, J. Evaluation of integrative clustering methods for the analysis of multi-omics data. **Briefings in Bioinformatics**, 2019.

CORCES, M. R.; GRANJA, J. M.; SHAMS, S.; LOUIE, B. H.; SEOANE, J. A.; ZHOU, W.; SILVA, T. C.; GROENEVELD, C.; WONG, C. K.; CHO, S. W.; et al. The chromatin accessibility landscape of primary human cancers. **Science**, v. 362, n. 420, 2018.

DELGADO, F. M.; GÓMEZ-VELA, F. Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. **Artificial Intelligence in Medicine**, 2018.

DUFFY, M.; HARBECK, N.; NAP, M.; MOLINA, R.; NICOLINI, A.; SENKUS, E.; CARDOSO, F. Clinical use of biomarkers in breast cancer: Updated guidelines from the European Group on Tumor Markers (EGTM). **European Journal of Cancer**, v. 75, p. 284–298, 2017. Pergamon.

EMMERT-STREIB, F.; DEHMER, M.; HAIBE-KAINS, B. Untangling statistical and biological models to understand network inference: the need for a genomics network ontology. **Frontiers in Genetics**, v. 5, p. 299, 2014. Frontiers.

FAITH, J. J.; HAYETE, B.; THADEN, J. T.; MOGNO, I.; WIERZBOWSKI, J.; COTTAREL, G.; KASIF, S.; COLLINS, J. J.; GARDNER, T. S. Large-Scale Mapping and Validation of Escherichia coli Transcriptional Regulation from a Compendium of Expression Profiles. (A. Levchenko, Org.) **PLoS Biology**, v. 5, n. 1, p. e8, 2007. Public Library of Science.

FARNHAM, P. J. Insights from genomic profiling of transcription factors. **Nature Reviews Genetics**, v. 10, 2009.

FLETCHER, M. N. C.; CASTRO, M. A. A.; WANG, X.; DE SANTIAGO, I.; O'REILLY, M.; CHIN, S.-F.; RUEDA, O. M.; CALDAS, C.; PONDER, B. A. J.; MARKOWETZ, F.; et al. Master regulators of FGFR2 signalling and breast cancer risk. **Nature Communications**, v. 4, 2013.

HANAHAHAN, D.; WEINBERG, R. A. Hallmarks of Cancer: The Next Generation. **Cell**, v. 144, n. 5, p. 646–674, 2011.

HELDRING, N.; PIKE, A.; ANDERSSON, S.; MATTHEWS, J.; CHENG, G.; HARTMAN, J.; TUJAGUE, M.; STRO" M, A.; STRO" M, S.; TREUTER, E.; et al. Estrogen Receptors: How Do

They Signal and What Are Their Targets., 2007.

JANGAL, M.; COUTURE, J.-P.; BIANCO, S.; MAGNANI, L.; MOHAMMED, H.; GÉVRY, N. The transcriptional co-repressor TLE3 suppresses basal signaling on a subset of estrogen receptor α target genes. **Nucleic acids research**, v. 42, n. 18, p. 11339–48, 2014. Oxford University Press.

KAMOUN, A.; REYNIES, A. DE; ALLORY, Y.; SJODAHL, G.; ROBERTSON, A. G.; SEILER, R.; HOADLEY, K.; AL-AHMADIE, H.; CHOI, W.; GROENEVELD, C. S.; et al. The consensus molecular classification of muscle-invasive bladder cancer. **bioRxiv**, p. 488460, 2018. Cold Spring Harbor Laboratory.

LAMB, J.; CRAWFORD, E. D.; PECK, D.; MODELL, J. W.; BLAT, I. C.; WROBEL, M. J.; LERNER, J.; BRUNET, J. P.; SUBRAMANIAN, A.; ROSS, K. N.; et al. The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. **Science**, v. 313, n. 5795, p. 1929–1935, 2006.

LEFEBVRE, C.; RAJBHANDARI, P.; ALVAREZ, M. J.; BANDARU, P.; LIM, W. K.; SATO, M.; WANG, K.; SUMAZIN, P.; KUSTAGI, M.; BISIKIRSKA, B. C.; et al. A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. **Molecular Systems Biology**, v. 377, 2010.

LIU, J.; LICHTENBERG, T.; HOADLEY, K. A.; POISSON, L. M.; LAZAR, A. J.; CHERNIACK, A. D.; KOVATICH, A. J.; BENZ, C. C.; LEVINE, D. A.; LEE, A. V.; et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. **Cell**, v. 173, n. 2, p. 400–416.e11, 2018.

LOCK, E. F.; DUNSON, D. B. Bayesian consensus clustering. **Bioinformatics**, v. 29, n. 20, p. 2610–2616, 2013. Oxford University Press.

MARGOLIN, A. A.; NEMENMAN, I.; BASSO, K.; WIGGINS, C.; STOLOVITZKY, G.; DALLA FAVERA, R.; CALIFANO, A. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. **BMC bioinformatics**, v. 7 Suppl 1, n. Suppl 1, p. S7, 2006. BioMed Central.

MATOS SIMOES, R. DE; DEHMER, M.; EMMERT-STREIB, F. Interfacing cellular networks of *S. cerevisiae* and *E. coli*: Connecting dynamic and genetic information. **BMC Genomics**,

v. 14, n. 1, p. 324, 2013. BioMed Central.

MAYRAN, A.; DROUIN, J. Pioneer transcription factors shape the epigenetic landscape. **Journal of Biological Chemistry**, p. jbc.R117.001232, 2018.

MENG, C.; HELM, D.; FREJNO, M.; KUSTER, B. moCluster: Identifying Joint Patterns Across Multiple Omics Data Sets. **Journal of Proteome Research**, v. 15, n. 3, p. 755–765, 2016. American Chemical Society.

MEYER, P. E.; KONTOS, K.; LAFITTE, F.; BONTEMPI, G. Information-Theoretic Inference of Large Transcriptional Regulatory Networks. **EURASIP Journal on Bioinformatics and Systems Biology**, v. 2007, p. 1–9, 2007. Hindawi Publishing Corp.

NADERI, A.; MEYER, M.; DOWHAN, D. H. Cross-regulation between FOXA1 and ErbB2 signaling in estrogen receptor-negative breast cancer. **Neoplasia (New York, N.Y.)**, v. 14, n. 4, p. 283–96, 2012. Neoplasia Press.

PEROU, C. M.; BØRRESEN-DALE, A.-L. Systems Biology and Genomics of Breast Cancer. **Cold Spring Harbor Perspectives in Biology**, v. 3, 2011.

ROBERTSON, A. G.; KIM, J.; AL-AHMADIE, H.; BELLMUNT, J.; GUO, G.; CHERNIACK, A. D.; HINOUE, T.; LAIRD, P. W.; HOADLEY, K. A.; AKBANI, R.; et al. Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. **Cell**, v. 171, n. 3, p. 540–556.e25, 2017. Cell Press.

SANDELIN, A.; CARNINCI, P.; LENHARD, B.; PONJAVIC, J.; HAYASHIZAKI, Y.; HUME, D. A. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. **Nature Reviews Genetics**, v. 8, n. 6, p. 424–436, 2007. Nature Publishing Group.

SHEN, R.; WANG, S.; MO, Q. Sparse Integrative Clustering of Multiple Omics Data Sets. **The annals of applied statistics**, v. 7, n. 1, p. 269–294, 2013. NIH Public Access.

SILVA, T. C.; COETZEE, S. G.; GULL, N.; YAO, L.; HAZELETT, D. J.; NOUSHMEHR, H.; LIN, D.-C.; BERMAN, B. P. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. (O. Stegle, Org.) **Bioinformatics**, 2018.

SORLIE, T.; TIBSHIRANI, R.; PARKER, J.; HASTIE, T.; MARRON, J. S.; NOBEL, A.; DENG, S.; JOHNSEN, H.; PESICH, R.; GEISLER, S.; et al. Repeated observation of breast tumor

subtypes in independent gene expression data sets. **Proceedings of the National Academy of Sciences of the United States of America**, v. 100, n. 14, p. 8418–23, 2003. National Academy of Sciences.

SPITZ, F.; M FURLONG, E. E. Transcription factors: from enhancer binding to developmental control. **Nature Reviews Genetics**, v. 13, 2012.

STEWART, B. W.; WILD, C.; INTERNATIONAL AGENCY FOR RESEARCH ON CANCER; WORLD HEALTH ORGANIZATION. **World cancer report 2014**. 2014.

SUBRAMANIAN, A.; TAMAYO, P.; MOOTHA, V. K.; MUKHERJEE, S.; EBERT, B. L.; GILLETTE, M. A.; PAULOVICH, A.; POMEROY, S. L.; GOLUB, T. R.; LANDER, E. S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 43, p. 15545–15550, 2005.

THEODOROU, V.; STARK, R.; MENON, S.; CARROLL, J. S. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. **Genome research**, v. 23, n. 1, p. 12–22, 2013. Cold Spring Harbor Laboratory Press.

VASKE, C. J.; BENZ, S. C.; SANBORN, J. Z.; EARL, D.; SZETO, C.; ZHU, J.; HAUSSLER, D.; STUART, J. M. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. **Bioinformatics**, v. 26, n. 12, p. i237–i245, 2010. Oxford University Press.

WERNER, H. M. J.; MILLS, G. B.; RAM, P. T. Cancer Systems Biology: a peek into the future of patient care? **Nature Reviews Clinical Oncology**, v. 11, n. 3, p. 167–176, 2014. Nature Publishing Group.

YAMAGUCHI, N.; NAKAYAMA, Y.; YAMAGUCHI, N. Down-regulation of Forkhead box protein A1 (FOXA1) leads to cancer stem cell-like properties in tamoxifen-resistant breast cancer cells through induction of interleukin-6. **Journal of Biological Chemistry**, v. 292, n. 20, p. 8136–8148, 2017.

YUAN, L.; GUO, L.-H.; YUAN, C.-A.; ZHANG, Y.-H.; HAN, K.; NANDI, A.; HONIG, B.; HUANG, D.-S. Integration of Multi-omics Data for Gene Regulatory Network Inference and Application to Breast Cancer. **IEEE/ACM Transactions on Computational Biology and**

Bioinformatics, p. 1–1, 2018.

ANEXOS

ANEXO I - THE CHROMATIN ACCESSIBILITY LANDSCAPE OF PRIMARY HUMAN
CANCERS

RESEARCH ARTICLE SUMMARY

CANCER

The chromatin accessibility landscape of primary human cancers

M. Ryan Corces*, Jeffrey M. Granja*, Shadi Shams, Bryan H. Louie, Jose A. Seoane, Wanding Zhou, Tiago C. Silva, Clarice Groeneveld, Christopher K. Wong, Seung Woo Cho, Ansuman T. Satpathy, Maxwell R. Mumbach, Katherine A. Hoadley, A. Gordon Robertson, Nathan C. Sheffield, Ina Felau, Mauro A. A. Castro, Benjamin P. Berman, Louis M. Staudt, Jean C. Zenklusen, Peter W. Laird, Christina Curtis, The Cancer Genome Atlas Analysis Network, William J. Greenleaf†, Howard Y. Chang†

INTRODUCTION: Cancer is one of the leading causes of death worldwide. Although the 2% of the human genome that encodes proteins has been extensively studied, much remains to be learned about the noncoding genome and gene regulation in cancer. Genes are turned on and off in the proper cell types and cell states by transcription factor (TF) proteins acting on DNA regulatory elements that are scattered over the vast noncoding genome and exert long-range influences. The Cancer Genome Atlas (TCGA) is a global consortium that aims to accelerate the understanding of the molecular basis of cancer. TCGA has systematically collected DNA mutation, methyl-

ation, RNA expression, and other comprehensive datasets from primary human cancer tissue. TCGA has served as an invaluable resource for the identification of genomic aberrations, altered transcriptional networks, and cancer subtypes. Nonetheless, the gene regulatory landscapes of these tumors have largely been inferred through indirect means.

RATIONALE: A hallmark of active DNA regulatory elements is chromatin accessibility. Eukaryotic genomes are compacted in chromatin, a complex of DNA and proteins, and only the active regulatory elements are accessible by the cell's machinery such as TFs. The assay for

transposase-accessible chromatin using sequencing (ATAC-seq) quantifies DNA accessibility through the use of transposase enzymes that insert sequencing adapters at these accessible chromatin sites. ATAC-seq enables the genome-wide profiling of TF binding events that orchestrate gene expression programs and give a cell its identity.

RESULTS: We generated high-quality ATAC-seq data in 410 tumor samples from TCGA, identifying diverse regulatory landscapes across 23 cancer types. These chromatin accessibility profiles identify cancer- and tissue-specific DNA regulatory elements that enable classification of

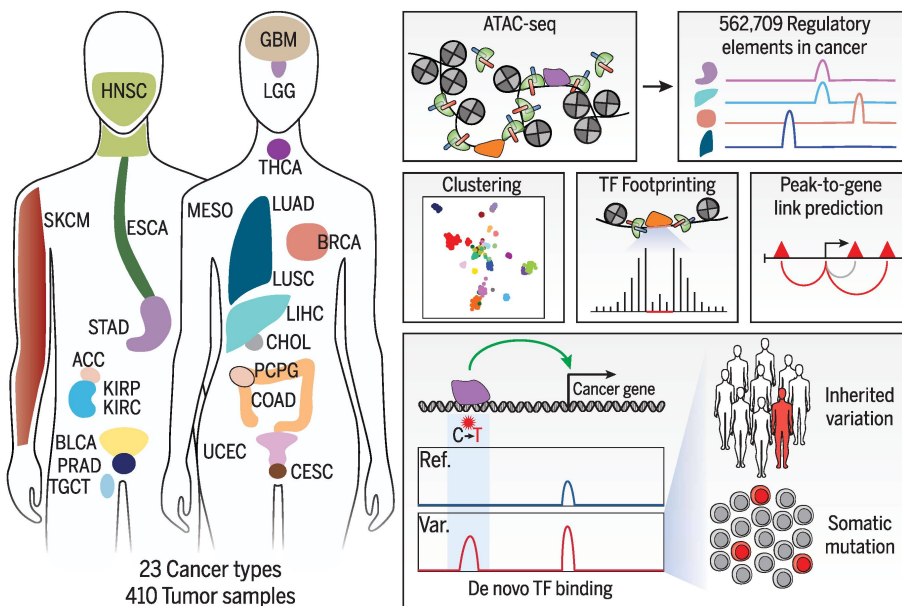
ON OUR WEBSITE

Read the full article at <http://dx.doi.org/10.1126/science.aav1898>

tumor subtypes with newly recognized prognostic importance. We identify distinct TF activities in cancer based on differences in the inferred patterns of TF-DNA interaction and gene

expression. Genome-wide correlation of gene expression and chromatin accessibility predicts tens of thousands of putative interactions between distal regulatory elements and gene promoters, including key oncogenes and targets in cancer immunotherapy, such as *MYC*, *SRC*, *BCL2*, and *PDL1*. Moreover, these regulatory interactions inform known genetic risk loci linked to cancer predisposition, nominating biochemical mechanisms and target genes for many cancer-linked genetic variants. Lastly, integration with mutation profiling by whole-genome sequencing identifies cancer-relevant noncoding mutations that are associated with altered gene expression. A single-base mutation located 12 kilobases upstream of the *FGD4* gene, a regulator of the actin cytoskeleton, generates a putative de novo binding site for an NKX TF and is associated with an increase in chromatin accessibility and a concomitant increase in *FGD4* gene expression.

CONCLUSION: The accessible genome of primary human cancers provides a wealth of information on the susceptibility, mechanisms, prognosis, and potential therapeutic strategies of diverse cancer types. Prediction of interactions between DNA regulatory elements and gene promoters sets the stage for future integrative gene regulatory network analyses. The discovery of hundreds of noncoding somatic mutations that exhibit allele-specific regulatory effects suggests a pervasive mechanism for cancer cells to manipulate gene expression and increase cellular fitness. These data may serve as a foundational resource for the cancer research community. ■



Cancer gene regulatory landscape. Chromatin accessibility profiling of 23 human cancer types (left) in 410 tumor samples from TCGA revealed 562,709 DNA regulatory elements. The activity of these DNA elements organized cancer subtypes, identified TF proteins and regulatory elements controlling cancer gene expression, and suggested molecular mechanisms for cancer-associated inherited variants and somatic mutations in the noncoding genome. See main article for abbreviations of cancer types. Ref., reference; Var., variant.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.

†Corresponding author. Email: howchang@stanford.edu (H.Y.C.); wjg@stanford.edu (W.J.G.)

Cite this article as M. R. Corces et al., *Science* 362, eaav1898 (2018). DOI: 10.1126/science.aav1898

RESEARCH ARTICLE

CANCER

The chromatin accessibility landscape of primary human cancers

M. Ryan Corces^{1*}, Jeffrey M. Granja^{1,2,3*}, Shadi Shams¹, Bryan H. Louie¹, Jose A. Seoane^{2,4,5}, Wanding Zhou⁶, Tiago C. Silva^{7,8}, Clarice Groeneveld⁹, Christopher K. Wong¹⁰, Seung Woo Cho¹, Ansuman T. Satpathy¹, Maxwell R. Mumbach^{1,2}, Katherine A. Hoadley¹¹, A. Gordon Robertson¹², Nathan C. Sheffield¹³, Ina Felau¹⁴, Mauro A. A. Castro⁹, Benjamin P. Berman⁷, Louis M. Staudt¹⁴, Jean C. Zenklusen¹⁴, Peter W. Laird⁶, Christina Curtis^{2,4,5}, The Cancer Genome Atlas Analysis Network[†], William J. Greenleaf^{1,2,3,15,16‡}, Howard Y. Chang^{1,2,17,18‡}

We present the genome-wide chromatin accessibility profiles of 410 tumor samples spanning 23 cancer types from The Cancer Genome Atlas (TCGA). We identify 562,709 transposase-accessible DNA elements that substantially extend the compendium of known cis-regulatory elements. Integration of ATAC-seq (the assay for transposase-accessible chromatin using sequencing) with TCGA multi-omic data identifies a large number of putative distal enhancers that distinguish molecular subtypes of cancers, uncovers specific driving transcription factors via protein-DNA footprints, and nominates long-range gene-regulatory interactions in cancer. These data reveal genetic risk loci of cancer predisposition as active DNA regulatory elements in cancer, identify gene-regulatory interactions underlying cancer immune evasion, and pinpoint noncoding mutations that drive enhancer activation and may affect patient survival. These results suggest a systematic approach to understanding the noncoding genome in cancer to advance diagnosis and therapy.

Cancer is a highly heterogeneous group of diseases, with each tumor type exhibiting distinct clinical features, patient outcomes, and therapeutic responses. The Cancer Genome Atlas (TCGA) was established to characterize this heterogeneity and understand the molecular underpinnings of cancer (1). Through large-scale genomic and molecular analyses, TCGA has revealed an exquisite diversity of genomic aberrations, altered transcriptional networks, and tumor subtypes that have engendered a more comprehensive understanding of disease etiologies and laid the foundations for new therapeutic and impactful clinical trials.

Work from TCGA and many others has demonstrated the importance of the epigenome to cancer initiation and progression (2). Profiling of cancer-specific coding mutations through whole-exome sequencing has identified prominent driver mutations in genes encoding chromatin remodel-

ing enzymes and modifiers of DNA methylation. These mutations drive alterations in the epigenome which, in turn, can establish the dysregulated cellular phenotypes that have become known as the hallmarks of cancer (3). Although many principles of chromatin regulation have been elucidated in cultured cancer cells, epigenomic studies of primary tumors are especially valuable, capturing the genuine ecosystem of heterotypic tumor and stromal cell interactions and the impacts of factors in the tumor microenvironment such as hypoxia, acidosis, and matrix stiffness (4). TCGA has carried out targeted DNA methylation profiling of more than 10,000 samples and, more recently, whole-genome bisulfite sequencing (WGBS) of 39 TCGA tumor samples (5). This data-rich resource has identified cancer-specific differentially methylated regions, providing an unprecedented view of epigenetic heterogeneity in cancer. In-

tegration of DNA methylation and additional TCGA data types has enabled the prediction of functional regulatory elements (6–8) and the identification of previously unknown cancer subtypes (9–13). Additional work has identified cancer-relevant variable enhancer loci by using histone modifications (14) and enhancer RNA sequencing (15). These studies represent, to date, the largest genome-wide epigenomic profiling efforts in primary human cancer samples.

Recently, the advent of the assay for transposase-accessible chromatin using sequencing (ATAC-seq) (16) has enabled the genome-wide profiling of chromatin accessibility in small quantities of frozen tissue (17). Because accessible chromatin is a hallmark of active DNA regulatory elements, ATAC-seq makes it possible to assess the gene regulatory landscape in primary human cancers. Combined with the richness of diverse, orthogonal data types in TCGA, the chromatin accessibility landscape in cancer provides a key link between inherited and somatic mutations, DNA methylation, long-range gene regulation, and, ultimately, gene expression changes that affect cancer prognosis and therapy.

Results

ATAC-seq in frozen human cancer samples is highly robust

We profiled the chromatin accessibility landscape for 23 types of primary human cancers, represented by 410 tumor samples derived from 404 donors from TCGA (protocol S1). These 23 cancer types are representative of the diversity of human cancers (Fig. 1A and data S1). From the 410 tumor samples, we generated technical replicates from 386 samples, yielding 796 genome-wide chromatin accessibility profiles (data S1). Given the size of this cohort, we first ensured that all generated ATAC-seq data could be uniquely mapped to the expected donor through comparison with single-nucleotide polymorphism (SNP) genotyping calls (fig. S1A). In all samples, the genotype from the ATAC-seq data generated in this study correlated most highly with previously published genotyping array data for the expected donor compared with that of all other 11,126 TCGA donors. All ATAC-seq data included in this study passed a minimum threshold of enrichment of signal over background (fig. S1, B to D, and data S1) with most samples showing a characteristic fragment size distribution with clear nucleosomal periodicity (fig. S1E). With this high-quality set of 410 tumor samples, we identified 562,709 reproducible (observed in more than one replicate) pan-cancer

¹Center for Personal Dynamic Regulomes, Stanford University, Stanford, CA 94305, USA. ²Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. ³Program in Biophysics, Stanford University School of Medicine, Stanford, CA 94305, USA. ⁴Department of Medicine, Stanford University School of Medicine, Stanford, CA 94305, USA. ⁵Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305, USA. ⁶Center for Epigenetics, Van Andel Research Institute, Grand Rapids, MI 49503, USA. ⁷Center for Bioinformatics and Functional Genomics, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA. ⁸Department of Genetics, Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, SP CEP 14.040-905, Brazil. ⁹Bioinformatics and Systems Biology Laboratory, Polytechnic Center, Federal University of Paraná, Curitiba, PR CEP 80.060-000, Brazil. ¹⁰Department of Biomolecular Engineering, Center for Biomolecular Sciences and Engineering, University of California–Santa Cruz, Santa Cruz, CA 95064, USA. ¹¹Department of Genetics, Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. ¹²Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, BC V5Z 4S6, Canada. ¹³Center for Public Health Genomics, University of Virginia, Charlottesville, VA 22908, USA. ¹⁴National Cancer Institute, NIH, Bethesda, MD 20892, USA. ¹⁵Department of Applied Physics, Stanford University, Stanford, CA 94025, USA. ¹⁶Chan Zuckerberg Biohub, San Francisco, CA 94158, USA. ¹⁷Program in Epithelial Biology, Stanford University, Stanford, CA 94305, USA. ¹⁸Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA.

*These authors contributed equally to this work. †The Cancer Genome Atlas Analysis Network collaborators and affiliations are listed in the supplementary materials.

‡Corresponding author. Email: howchang@stanford.edu (H.Y.C.); wjg@stanford.edu (W.J.G.)

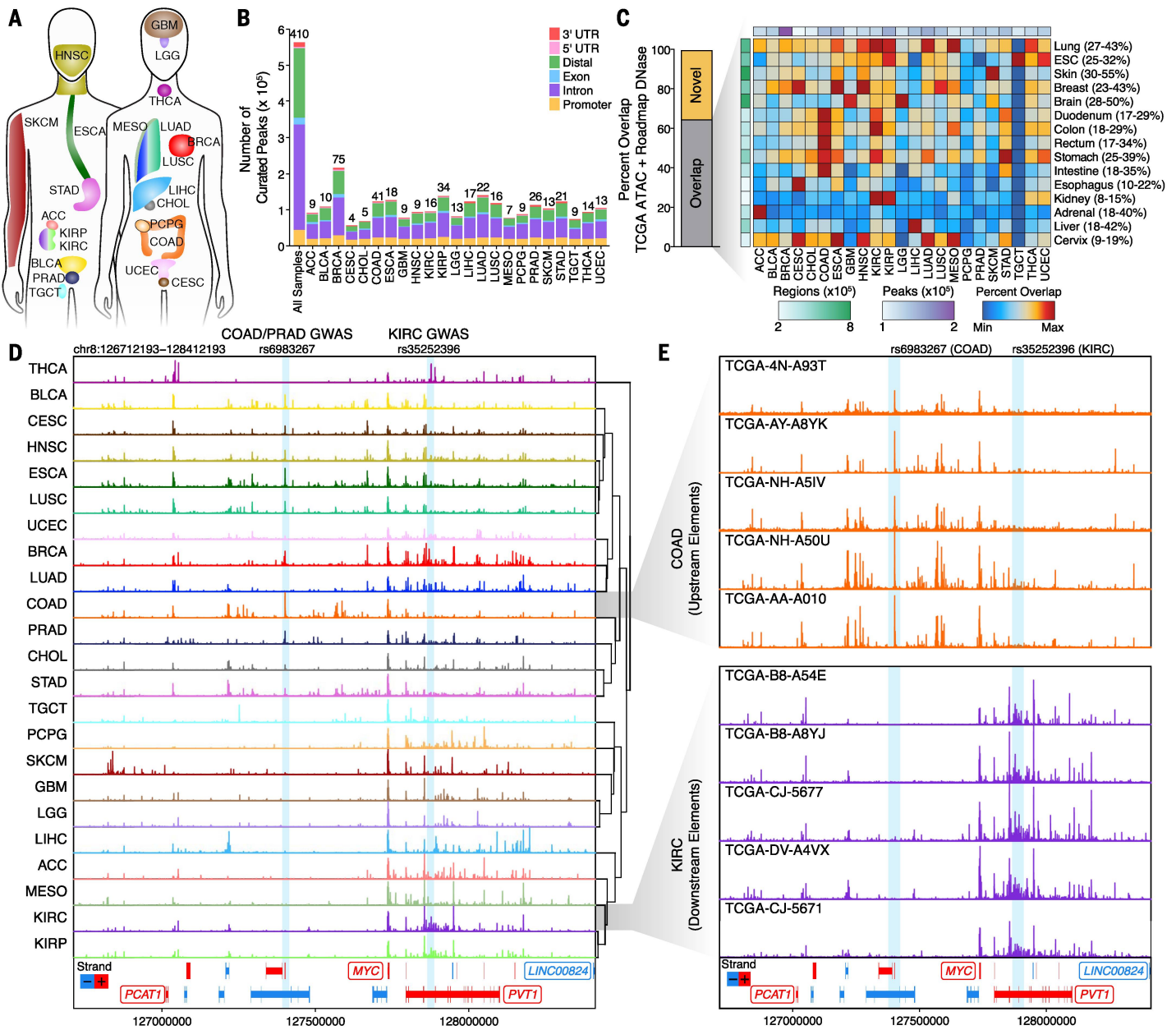


Fig. 1. Pan-cancer ATAC-seq of TCGA samples identifies diverse regulatory landscapes. (A) Diagram of the 23 cancer types profiled in this study. Colors are kept consistent throughout figures. (B) Pan-cancer peak calls from ATAC-seq data. Peak calls from each cancer type are shown individually in addition to the 562,709 peaks that represent the pan-cancer merged peak set. Color indicates the type of genomic region overlapped by the peak. The numbers shown above each bar represent the number of samples profiled for each cancer type. UTR, untranslated region. (C) Overlap of cancer type-specific ATAC-seq peaks with Roadmap DNase-seq peaks from various tissues and cell types. Left: The percent of ATAC-seq peaks that are overlapped by one or more Roadmap peaks. Right: A heatmap of the percent overlap observed for each ATAC-seq peak set within the Roadmap DNase-seq peak set. Colors are scaled according to the minimum and maximum overlaps, which are indicated numerically to the right of the DNase-seq peak set names. The total number of ATAC-seq peaks (white to purple) or Roadmap DNase-seq regions (white to green) are shown colorimetrically. (D) Normalized ATAC-seq sequencing tracks of all 23 cancer types at the MYC locus. Each track represents the average accessibility per 100-bp bin across all

replicates. Known GWAS SNPs rs6983267 (COAD, PRAD) and rs35252396 (KIRC) are highlighted with light blue shading. Region shown represents chromosome 8 (chr8):126712193 to 128412193. (E) Normalized ATAC-seq sequencing tracks of five different COAD samples (top, orange) and KIRC samples (bottom, purple) shown across the same MYC locus as in Fig. 1D. Known GWAS SNPs rs6983267 (COAD, PRAD) and rs35252396 (KIRC) are highlighted with light blue shading. Region shown represents chr8:126712193 to 128412193. ACC, adrenocortical carcinoma; BLCA, bladder urothelial carcinoma; BRCA, breast invasive carcinoma; CESC, cervical squamous cell carcinoma; CHOL, cholangiocarcinoma; COAD, colon adenocarcinoma; ESCA, esophageal carcinoma; GBM, glioblastoma multiforme; HNSC, head and neck squamous cell carcinoma; KIRC, kidney renal clear cell carcinoma; KIRP, kidney renal papillary cell carcinoma; LGG, low grade glioma; LIHC, liver hepatocellular carcinoma; LUAD, lung adenocarcinoma; LUSC, lung squamous cell carcinoma; MESO, mesothelioma; PCPG, pheochromocytoma and paraganglioma; PRAD, prostate adenocarcinoma; SKCM, skin cutaneous melanoma; STAD, stomach adenocarcinoma; TGCT, testicular germ cell tumors; THCA, thyroid carcinoma; UCEC, uterine corpus endometrial carcinoma.

peaks of chromatin accessibility (Fig. 1B and data S2). These peaks were identified using a normalized peak score metric to enable direct comparison of peaks across samples of unequal sequencing depth, with each cancer type having an average of 105,585 peaks (range 56,125 to 215,978; Fig. 1B and fig. S1F; see methods). Reproducibility within the pan-cancer peak set was high for technical replicates (different nuclei from the same tumor sample; fig. S1, G and H), intratumor replicates (different samples from the same tumor; fig. S1I), and intertumor replicates (tumor samples from different donors; fig. S1, J and K).

Cancer chromatin accessibility extends the dictionary of DNA regulatory elements

The pan-cancer and cancer type-specific peak sets generated in this study enabled quantification of the number of DNA regulatory elements identified. To do this, we compared the regions defined by our pan-cancer and cancer type-specific peak sets to the regions defined by the Roadmap Epigenomics Project deoxyribonuclease I hypersensitive sites sequencing (DNase-seq) studies (18), finding a median of 34.4% overlap between the cancer type-specific peak sets and the various Roadmap tissue-type peak sets, with the strongest overlap occurring in the expected combinations (Fig. 1C and data S3). In total, about 65% of the pan-cancer peaks identified in this study had overlap with previously observed regulatory elements, highlighting both the consistency of our results with published datasets and the large number of additional putative regulatory elements observed in this study (Fig. 1C). Given the extensive coverage of Roadmap DNase-seq studies in healthy tissues, our results suggested that the disease context of cancer unveils the activity of additional DNA regulatory elements. Moreover, overlap of the ATAC-seq-defined DNA regulatory elements with chromatin immunoprecipitation sequencing (ChIP-seq)-defined ChromHMM regulatory states shows a strong enrichment of accessible chromatin sites in promoter and enhancer regions, as expected (fig. S1L). Although we profiled many samples in some cancer types [i.e., breast invasive carcinoma (BRCA), 75 tumor samples], we profiled fewer samples in multiple other cancer types (i.e., cervical squamous cell carcinoma, four tumor samples) (Fig. 1B). By estimating the number of unique peaks added with each additional sample, we found that cancer types have an estimated average of 169,822 total peaks (range 97,995 to 309,313) at saturation (fig. S1, M and N, and data S3), suggesting that profiling of additional samples of each cancer type would further expand the repertoire of regulatory elements.

Noncoding DNA elements reveal distinct cancer gene regulation and genetic risks

The *MYC* proto-oncogene locus provides a prime illustration of the diversity of the chromatin accessibility landscape across cancer types. *MYC* is embedded in a region with multiple DNA

regulatory elements and noncoding transcripts that regulate *MYC* in a tissue-specific fashion (19). We observed sufficient diversity in the chromatin accessibility landscape of the *MYC* locus to enable clustering of cancer types into two primary categories: (i) cancer types with extensive chromatin accessibility at 5' and 3' DNA elements, such as colon adenocarcinoma (COAD), and (ii) cancer types with chromatin accessibility primarily at 3' regulatory elements, such as kidney renal clear cell carcinoma (KIRC) (Fig. 1D). This trend is consistent across different samples of the same cancer type, as shown for COAD and KIRC (Fig. 1E) and is similar to the regulation observed in the *HOXD* locus (20).

Genome-wide association studies (GWAS) have identified numerous inherited risk loci for cancer susceptibility. However, many of these SNPs reside in the noncoding genome within known DNA regulatory elements. In the *MYC* locus, we identify known sites of chromatin accessibility, including peaks surrounding functionally validated GWAS cancer susceptibility SNPs (rs6983267 and rs35252396; Fig. 1, D and E). SNP rs6983267 is associated with increased susceptibility to colon adenocarcinoma and prostate adenocarcinoma (PRAD) (21–23), consistent with the presence of focal chromatin accessibility in these cancer types. However, SNP rs6983267 has not been previously associated with breast cancer or any squamous tumor types, which also have strong chromatin accessibility at this regulatory element in our ATAC-seq data (Fig. 1D). Similarly, SNP rs35252396 has been associated with KIRC and, in our data, shows strong accessibility in samples from kidney cancer types as well as breast and thyroid carcinoma, suggesting a potential role for these SNPs in previously unappreciated cancer contexts.

To visualize global patterns from our diverse ATAC-seq datasets, we performed Pearson correlation hierarchical clustering on distal and promoter elements (Fig. 2A). We found that distal elements exhibited a greater specificity and wider dynamic range of activity in association with cancer types, whereas promoter element accessibility was less cancer type-specific and showed similar patterns of correlation to global gene expression, as measured by RNA-seq (Fig. 2A). This functional specificity of distal regulatory elements was also previously observed in healthy tissues and in development (24, 25). Using *t*-distributed stochastic neighbor embedding (26) (*t*-SNE; Fig. 2B) and density clustering (27) (fig. S2A), we identified 18 distinct clusters, which we labeled based on the observed cancer-type enrichment (fig. S2B and data S3). We found strong concordance between this ATAC-seq-based clustering and the published multiomic iCluster scheme using TCGA mRNA-seq, microRNA (miRNA)-seq, DNA methylation, reverse-phase protein array (RPPA), and DNA copy number data (28) (Fig. 2, C and D). Comparing this clustering scheme to other TCGA-based clustering schemes, we observed the strongest concordance of our ATAC-seq clustering scheme with mRNA and cancer type (Fig. 2E).

This is consistent with the connection of chromatin accessibility to transcriptional output and the observation that ATAC-seq is strongly cell type-specific. Multiple observations can be made from these clusters: (i) Some cancer types split into two distinct clusters such as breast cancer (i.e., basal and nonbasal) and esophageal cancer (i.e., squamous and adenocarcinoma), (ii) cancer samples derived from the same tissue type often group together [i.e., kidney renal papillary cell carcinoma (KIRP) and KIRC], and (iii) some cancers group together across tissues as observed for squamous cell types (Fig. 3A and fig. S2B).

Cluster-specific regulatory landscapes identify patterns of transcription factor usage and DNA hypomethylation

Grouping of samples into defined clusters enables the determination of patterns in chromatin accessibility that are unique to each cluster. Using a framework that we term “distal binarization,” we identified the distal regulatory elements that are accessible only in a single cluster or small group of clusters (Fig. 3B, fig. S2C, and data S4). Of the 516,927 pan-cancer distal elements, 203,260 were found to be highly accessible in a single cluster or group of clusters (up to four clusters). These cluster-specific peak sets are enriched for motifs of transcription factors (TFs) with correlated gene expression that are known to be important for cancer and tissue identity (Fig. 3C, fig. S2D, and data S4). These include the androgen receptor (AR) in prostate cancer, forkhead box A1 (FOXA1) in nonbasal breast cancer, and melanogenesis-associated transcription factor (MITF) in melanoma. Moreover, these cluster-specific peak sets are enriched for known GWAS SNPs that are associated with cancers of the corresponding type (fig. S2E and data S5), highlighting that cancer-related GWAS SNPs tend to be located within or near cancer type-specific regulatory elements. The concordance of GWAS risk loci and cancer chromatin state has often been evaluated using cancer cell lines in the past, and our work provides a foundational map to evaluate noncoding GWAS SNPs in primary human cancers.

Consistent with published reports (12, 18, 29, 30), the degree of DNA methylation was anticorrelated with chromatin accessibility at regulatory elements, and regions lacking chromatin accessibility were more frequently methylated (fig. S2F). In particular, cluster-specific peak sets are hypomethylated in the relevant cancer types, though frequently methylated in other cancer types that lack accessibility in those peaks (fig. S2G). Consistent with these observations, which are based on DNA methylation array data, we see a strong depletion of DNA methylation at the center of both distal peaks and promoter peaks in a single patient profiled by WGBS (fig. S2H) (5). In our analysis of methylation levels within cluster-specific peak sets, we also identified a subgroup of brain cancers that exhibits DNA hypermethylation of peaks specific to nonbrain cancers (fig. S2G), likely caused by mutations in genes

that affect DNA methylation, such as isocitrate dehydrogenase 1 (*IDH1*) (fig. S3A). Similarly, we found that the subset of testicular germ cell tumors that are seminomas show a pattern of genome-wide DNA hypomethylation, consistent with a published report (31) (fig. S3B). Thus, a small number of TFs dominate the cis-regulatory landscape in each cancer type. These TFs are often the known key drivers of the respective

cancer or tissue type, and TF occupancy is associated with, and possibly causes, DNA hypomethylation of the corresponding DNA elements in cancer.

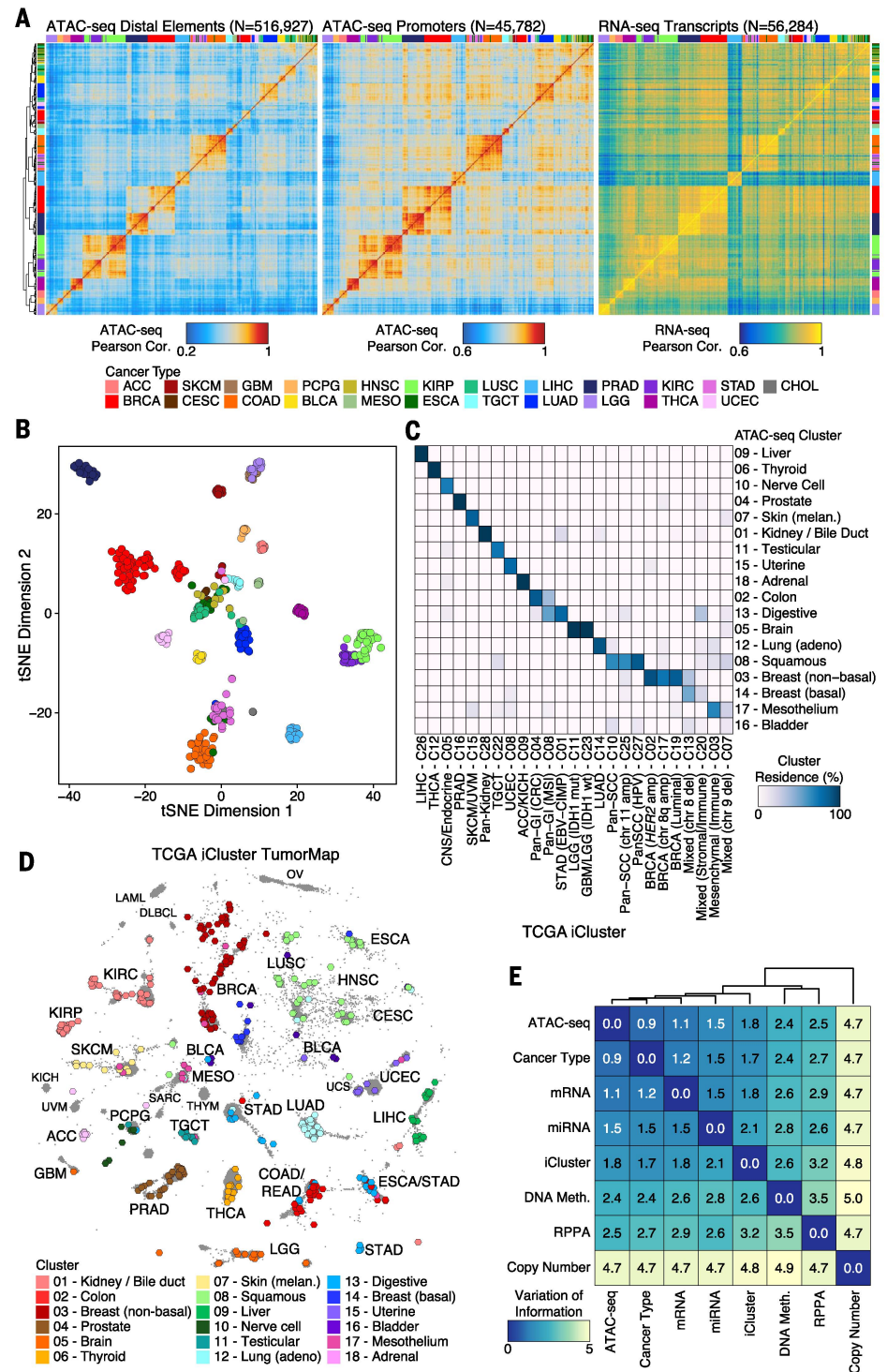
De novo identification of cancer subtypes from ATAC-seq data

Given the richness of the chromatin accessibility landscape, we explored the capacity of

ATAC-seq data to define molecular subtypes of cancer de novo. This analysis was limited to cancer types with sufficient available donors: BRCA (*N* = 74), PRAD (*N* = 26), and KIRP (*N* = 34). In KIRP, a gap statistic identified three distinct subgroups that are clearly separable by the first two principal components (Fig. 3D). The smallest of these subgroups contains four donors with very clear differences in ATAC-seq

Fig. 2. Chromatin accessibility profiles reveal distinct molecular subtypes of cancers.

(A) Pearson correlation heatmaps of ATAC-seq distal elements (left), ATAC-seq promoters (middle), and RNA-seq of all genes (right). Clustering orientation is dictated by the ATAC-seq distal element accessibility, and all other heatmaps use this same clustering orientation. Color scale values vary between heatmaps. Promoter peaks are defined as occurring between -1000 and +100 bp of a transcriptional start site. Distal peaks are all nonpromoter peaks. The total number of features (*N*) used for correlation is indicated above each Pearson correlation heatmap. **(B)** Unsupervised t-SNE on the top 50 principal components for the 250,000 most variable peaks across all cancer types. Each dot represents the merge of all technical replicates from a given sample. Color represents the cancer type shown above the plot. **(C)** Cluster residence heatmap showing the percent of each TCGA iCluster that overlaps with each ATAC-seq-based cluster. **(D)** ATAC-seq t-SNE clusters shown on the PanCanAtlas iCluster TumorMap. Each hexagon represents a cancer patient sample, and the positions of the hexagons are computed from the similarity of samples in the iCluster latent space. The color and larger size of the hexagon indicates the ATAC-seq cluster assignment. Samples that were not included in the ATAC-seq analysis are represented by smaller gray hexagons. The text labels indicate the cancer disease type. **(E)** Variation of information analysis of clustering schemes derived by using various data types from TCGA.



Downloaded from <http://science.sciencemag.org/> on January 14, 2019

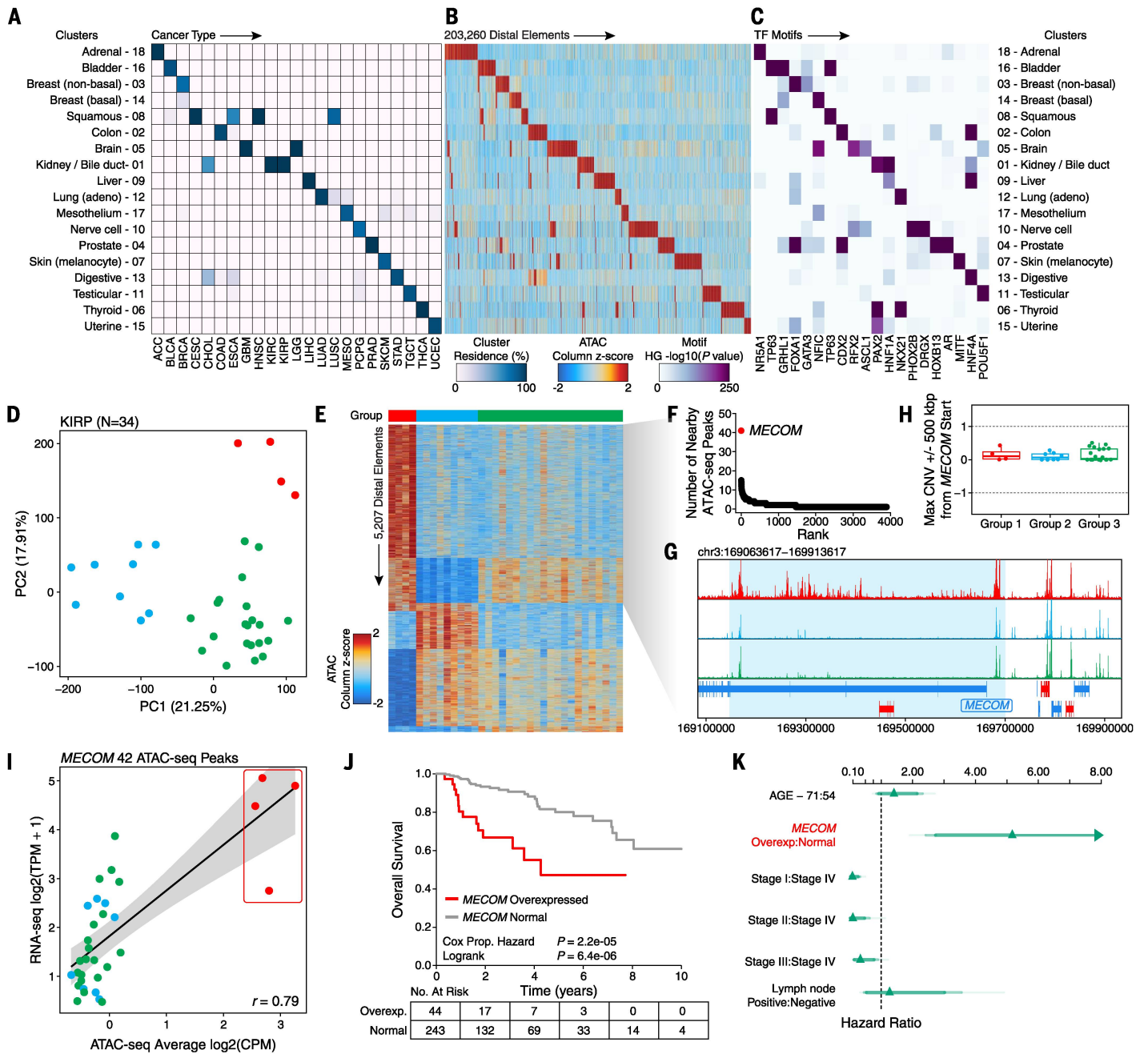


Fig. 3. ATAC-seq clusters cancer samples to show cancer- and tissue-specific drivers. (A) Cluster residence heatmap showing the percent of samples from a given cancer type that reside within each of the 18 annotated ATAC-seq clusters. (B) Heatmap showing the ATAC-seq accessibility at distal elements ($N = 203,260$) identified to be cluster-specific by distal binarization. (C) Enrichment of TF motifs in peak sets identified in Fig. 3B. Enrichment is determined by a hypergeometric (HG) test $-\log_{10}(P$ value) of the motif's representation within the cluster-specific peaks compared to the pan-cancer peak set. Transcription factors shown represent a manually trimmed set of factors whose expression is highly correlated ($r > 0.4$) with the accessibility of the corresponding motif. Color represents the $-\log_{10}(P$ value) of the hypergeometric test. (D) Principal component analysis of the top 25,000 distal ATAC-seq peaks within the KIRP cohort ($N = 34$ samples). Each dot represents an individual sample. The color of the dots represents K -means clustering ($K = 3$ by gap statistic). (E) Distal binarization analysis based on the three K -means-defined groups identified and shown (by color) in

Fig. 3D. (F) Dot plot showing the number of nearby ATAC-seq peaks per gene from the group 1 distal binarization. Each dot represents a different gene. The *MECOM* gene (also called *EVII*) is highlighted in red. (G) Normalized average sequencing tracks of K -means-defined groups 1, 2, and 3 at the *MECOM* locus. Peaks specific to group 1 are highlighted by light blue shading. (H) DNA copy number data at the *MECOM* locus in the three K -means-defined groups. Each dot represents an individual sample. CNV, copy number variation. (I) Average chromatin accessibility at peaks near the *MECOM* gene ($N = 42$ peaks) and RNA-seq gene expression of *MECOM* in KIRP samples ($N = 34$ samples). Each dot represents an individual donor. Dots are colored according to the clustering group colors shown in Fig. 3D. CPM, counts per million. (J) Kaplan-Meier analysis of overall survival of all KIRP donors in TCGA ($N = 287$) stratified by *MECOM* overexpressed ($N = 44$) and normal *MECOM* expression ($N = 243$). (K) Hazard plot of risk of dying from KIRP based on multiple covariates, including *MECOM* expression (hazard ratio = 5.2, 95% confidence interval = 2.4 to 11.0). Lines represent 95% confidence intervals.

accessibility identified by distal binarization (red coloring in Fig. 3E). Within the set of regulatory elements that are specific to this subgroup, we found 42 ATAC-seq peaks near the MDS1 and EVI1 complex locus (*MECOM*) gene (Fig. 3, F and G). Notably, the high chromatin accessibility of these *MECOM* peaks is not related to copy number amplification, as determined by DNA copy number array data (Fig. 3H). The expression of the *MECOM* gene is highly correlated with the mean ATAC-seq accessibility at these 42 ATAC-seq peaks [correlation coefficient (r) = 0.79, Fig. 3I]. Additionally, overexpression of *MECOM* is significantly associated with poorer overall survival across all available KIRP data from TCGA ($P = 2.2 \times 10^{-5}$, Cox proportional hazard test, Fig. 3J) with a hazard ratio of 5.2 (95% confidence interval = 2.4 to 11.0). This association is more substantial than lymph node status or patient age and is independent of cancer stage (Fig. 3K), indicating a potential prognostic role for these findings. Importantly, *MECOM* overexpression is not readily explained by any previously identified subgroups of KIRP, including subgroups with a CpG island methylator phenotype or mutations in the gene encoding fumarate hydratase, which have also been shown to confer poor overall survival (13). These results suggest that *MECOM* activation in KIRP identifies a previously unappreciated subgroup of patients with adverse outcomes, a finding that was uncovered by notable changes in the chromatin accessibility landscape of these samples.

Similarly, we found multiple distinct subgroups of PRAD and BRCA based on *K*-means clustering of the top 25,000 variable distal ATAC-seq peaks (fig. S3, C and D). In PRAD, these include subgroups driven by activity of AR, tumor protein P63 (*TP63*), and forkhead box-family TFs (fig. S3C). From an unsupervised analysis of breast cancer, we identified motifs of known TF drivers of luminal subtype identity, such as GATA binding protein 3 (*GATA3*) and FOXA1, as being enriched in the peak clusters specific to a subset of luminal samples (clusters 3 and 4, fig. S3D). We also identified a potential role for grainyhead-like (*GRHL*) TF motifs in basal breast cancer (32) (cluster 1, fig. S3D) and an overlapping role for nuclear factor I (*NFI*) in both basal and luminal A breast cancer (cluster 2, fig. S3D). Additionally, ATAC-seq data can be used to identify regions of copy number amplification de novo (33), enabling the classification of *HER2*-amplified cases of breast cancer (fig. S3, E to G).

Footprinting analysis defines TF activities in cancer

The high sequencing depth of the ATAC-seq data generated in this study (median of 56.7 million unique reads per technical replicate) enabled the profiling of TF occupancy at base-pair resolution through TF footprinting. TF binding to DNA protects the protein-DNA binding site from transposition while the displacement or depletion of one or more nucleosomes creates high DNA accessibility in the immediate flanking se-

quence. Collectively, these phenomena are referred to as the TF footprint. To characterize TF footprints, we adapted a recent approach (34) that quantifies the “flanking accessibility,” a measure of the accessibility of the DNA adjacent to a TF motif, and “footprint depth,” a measure of the relative protection of the motif site from transposition (Fig. 4A and data S6). To calculate these variables, we aggregated all insertions relative to the TF motif center, genome-wide (fig. S4A). To attempt to account for known Tn5 transposase insertion bias, we computed the hexamer frequency centered at Tn5 insertions and normalized for the expected bias at each position relative to the motif center (34) (see methods for potential limitations). Depending on the binding properties of a TF and its ability to affect local chromatin accessibility, changes in these properties would be detectable through this approach genome-wide (fig. S4, B and C). ChromVAR (35), a similar genome-wide approach which assesses the ability of a TF to affect flanking accessibility, identified a highly overlapping list of TFs (fig. S4D).

To uncover transcriptionally driven TF binding patterns, we correlated the RNA-seq gene expression of a given TF to its corresponding footprint depth and estimated flanking accessibility (data S6). A factor whose expression is sufficient to generate robust DNA binding would have a footprint depth and flanking accessibility that are significantly correlated to its gene expression [false discovery rate (FDR) < 0.1, purple dots in Fig. 4B], such as TP63 (Fig. 4, C and D) or NK2 homeobox 1 (*NKX2-1*) (Fig. 4, E and F). Increases in flanking accessibility and decreases in footprint depth are likewise accompanied by decreases in methylation (bottom of Fig. 4, D and F), consistent with the hypothesis that methylated DNA is less likely to be bound by TFs (36). Although footprint depth and flanking accessibility are often correlated, their divergence can suggest the modes of TF-DNA interaction. For example, factors whose expression is sufficient to cause opening of chromatin around the motif site but not to protect the motif site from transposition would be expected to only exhibit a significant correlation between gene expression and flanking accessibility (blue dots in Fig. 4B). This pattern of correlation could be caused by effects such as rapid TF off rates or low occupancy (fig. S4, E and F). Conversely, a small number of TFs have expression that is only significantly correlated with footprint depth (red dots in Fig. 4B). Though likewise rare, we also identified potential negative regulators whose expression is inversely correlated to gain of flanking accessibility and loss of footprint depth, such as the cut-like homeobox 1 (*CUX1*) TF (37) (Fig. 4B and fig. S4, G and H). This is the expected behavior of repressive TFs that bind DNA and lead to compaction of the neighboring sequence. These results predicted dozens of positive and negative regulators whose expression is strongly correlated with chromatin accessibility patterns near to their corresponding motif (fig. S4I and data S6). Overall, our

footprinting analysis identified putative TFs with activities correlated with gene expression.

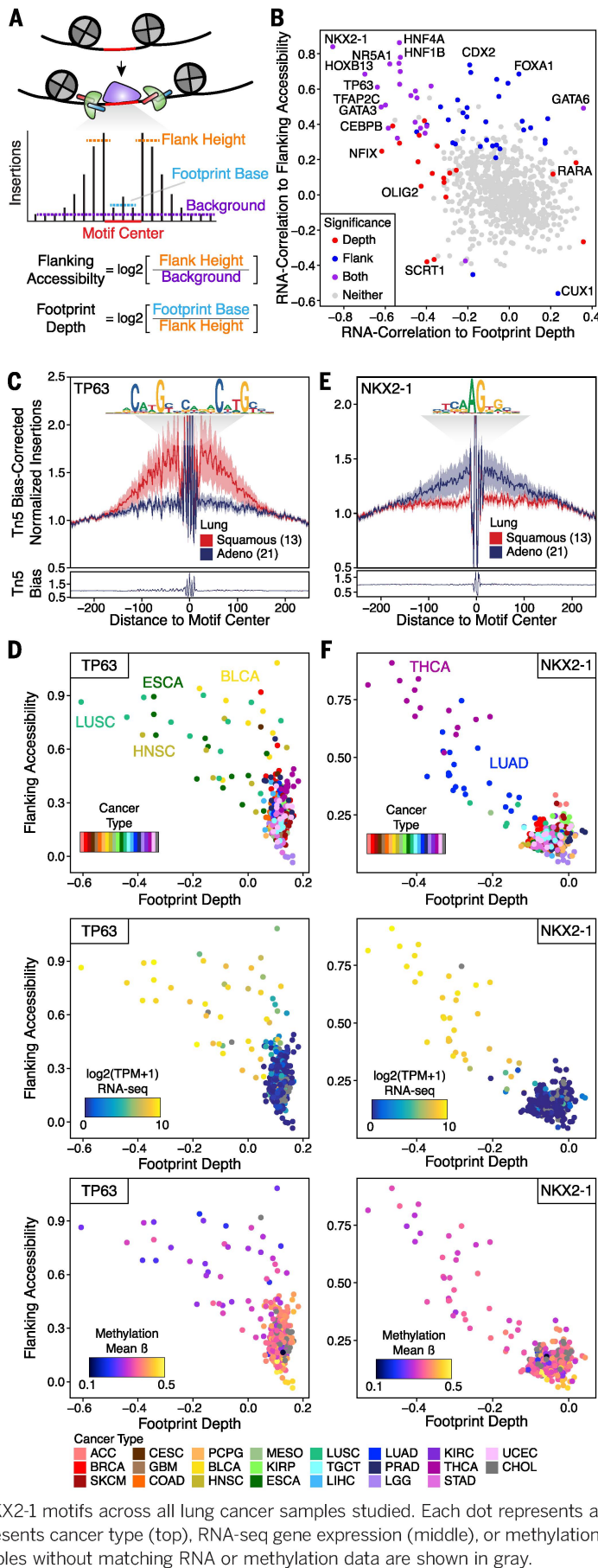
Linking of DNA regulatory elements to genes predicts interactions relevant to cancer biology

The breadth and depth of this sequencing study enabled a robust association of ATAC-seq peaks with the genes that they are predicted to regulate. To do this, we implemented a strategy based on the correlation of ATAC-seq accessibility and gene expression across all samples (Fig. 5A, $N = 373$ with matched RNA-seq and ATAC-seq). Because promoter capture Hi-C data suggested that >75% of three-dimensional (3D) promoter-based interactions occur within a 500-kilobase pair (kbp) distance (38), we restricted the length scale of this analysis to 500 kbp to avoid spurious predictions. Using a conservative FDR cutoff of 0.01, we identified 81,323 unique links between distal ATAC-seq peaks and genes (Fig. 5B and data S7). Some of these links are driven by correlation across many cancer types (Fig. 5, C to E), whereas 70% are strongly driven by one cluster (Fig. 5F and data S7). To derive a final list of peak-to-gene links (Fig. 5B), putative links were filtered against (i) links whose correlation is strongly driven by DNA copy number amplification (“CNA”; fig. S5, A and B), (ii) regions with broad and high local correlation (“diffuse”; fig. S5, B and C), and (iii) links involving an ATAC-seq peak that overlaps the promoter of any gene (Fig. 5G). As expected, the histogram of distances between a peak and its target gene decays sharply with distance (39) (Fig. 5H). The expression of most genes is correlated with the activity of fewer than five different peaks (Fig. 5I), whereas most peaks are predicted to interact with a single gene (Fig. 5J). Additionally, this analysis found that only 24% of predicted links occur between an ATAC-seq peak and the nearest gene, indicating that the majority of predicted interactions skip over one or more genes and would not be possible to predict from primary sequence alone (Fig. 5K). In total, we predicted at least one peak-to-gene link for 8552 protein-coding genes, accounting for nearly half of all protein-coding genes in the human genome, including 48% of the curated Catalogue of Somatic Mutations in Cancer (COSMIC) cancer-relevant genes (data S7).

In addition to predicting peak-to-gene links across cancer types, we also predicted peak-to-gene links within breast cancer ($N = 74$ donors), identifying 9711 unique peak-to-gene links (fig. S5D and data S7). Of these links, 36% were also identified in our analysis of all cancer types (fig. S5E). Particularly important in these BRCA-specific links was the contribution of recurrent DNA CNA as a strong driver for spurious peak-to-gene correlation (Fig. 5G). These false-positive associations were removed through the use of published TCGA DNA copy number array data and a local correlation correction model, as mentioned above (see methods). The final predicted BRCA-specific links follow a similar distance

Fig. 4. Footprinting analysis identifies distinct TF activities in cancer.

(A) Schematic illustrating the dynamics of TF binding (purple) and Tn5 insertion (green). **(B)** Classification of TFs by the correlation of their RNA expression to the footprint depth and flanking accessibility of their motifs. Color represents whether the depth (red), flank (blue), or both (purple) are significantly correlated to TF expression below an FDR cutoff of 0.1. Each dot represents an individual deduplicated TF motif (see methods). **(C)** TF footprinting of the TP63 motif (CIS-BP M2321_1.02) in lung cancer samples from the squamous (cluster 8) or adenocarcinoma (cluster 12) subtype. The Tn5 insertion bias track of TP63 motifs is shown below. **(D)** Dot plots showing the footprint depth and flanking accessibility of TP63 motifs across all lung cancer samples studied. Each dot represents a unique sample. Color represents cancer type (top), RNA-seq gene expression (middle), or methylation beta value (bottom). Samples without matching RNA or methylation data are shown in gray. **(E)** TF footprinting of the NKX2-1 motif (CIS-BP M6374_1.02) in lung cancer samples from the squamous (cluster 8) and adenocarcinoma (cluster 12) subtype. The Tn5 insertion bias track of NKX2-1 motifs is shown below. **(F)** Dot plots showing the footprint depth and flanking accessibility of NKX2-1 motifs across all lung cancer samples studied. Each dot represents a unique sample. Color represents cancer type (top), RNA-seq gene expression (middle), or methylation beta value (bottom). Samples without matching RNA or methylation data are shown in gray.



distribution and peak-to-gene linking specificity as observed in the pan-cancer predicted links (fig. S5, F to I).

Many of these predicted peak-to-gene links occur in clusters where multiple nearby peaks are predicted to be linked to the same gene, indicating that these clusters of peak-to-gene links may function as part of a single regulatory unit or enhancer. Extending the width of the linked ATAC-seq peaks to 1500 bp allows for joining of these peaks into defined merged putative enhancer units (fig. S5J). This resulted in a total of 58,092 pan-cancer and 7622 BRCA-specific enhancer-to-gene links (data S7).

Validation and utility of predicted links between distal elements and genes

To verify a regulatory interaction for the predicted peak-to-gene links, we used a CRISPR interference (CRISPRi) (40) strategy using a catalytically dead Cas9 (dCas9) fused to a Kruppel-associated box (KRAB) domain, which mediates focal heterochromatin formation and functional silencing of noncoding DNA regulatory elements (Fig. 6A). In this way, targeting the distal peak region of a predicted peak-to-gene link would be expected to cause a decrease in the expression of the linked gene, located tens to hundreds of kilobases away. CRISPRi of a predicted distal regulatory element linked to *BCL2* (164 kbp, Fig. 5C) led to a significant reduction in *BCL2* gene expression in the luminal-like breast cancer MCF7 cell line but not in the basal-like MDA-MB-231 cell line (Fig. 6B), consistent with the role of *BCL2* as a luminal-specific survival factor (41). Similarly, CRISPRi of a distal regulatory element linked to the *SRC* oncogene (-49 kbp, Fig. 5D) led to a significant reduction in gene expression in both MCF7 cells and MDA-MB-231 cells (Fig. 6B). On a genome-wide scale, the predicted BRCA-specific peak-to-gene links show a strong enrichment in 3D chromosome conformation data from MDA-MB-231 cells (42), providing further support for our link prediction strategy (Fig. 6C). Moreover, we found that, of the peak-to-gene links predicted from BRCA ATAC-seq data that are also associated with a DNA methylation array CpG probe, 35% overlap with links predicted jointly from DNA methylation array and RNA-seq data in an ELMER analysis (8, 43) of the complete TCGA BRCA dataset ($N = 858$ tumors) ($P < 0.001$; Fig. 6D, fig. S6A, and data S8). These overlaps contain many luminal-specific and basal-specific links (fig. S6A), with a clear delineation between luminal (fig. S6B) and basal (fig. S6C) breast cancer samples. Integrating WGBS and ATAC-seq demonstrated the dynamics of methylation and chromatin accessibility and the overlap of predicted interactions at the non-basal *FOXA1* and basal forkhead box C1 (*FOXC1*) loci (fig. S6, D and E).

Similarly, previous work has leveraged TCGA RNA-seq data to infer transcriptional networks that consist of regulons, each of which is based on a TF regulator and its associated positive and negative target genes (fig. S7A) (44). For each

regulon, every donor in the cohort can be assigned a positive, undefined, or negative regulon activity as measured by a differential enrichment score (dES) (45). Certain patterns of chromatin accessibility are expected on the basis of the target gene set and dES status of the donor (fig. S7B). For example, in donors with positive dES, chromatin at sites linked to positive target genes should be more accessible, whereas chromatin at sites linked to negative targets should be less accessible (fig. S7B). Examination of the estrogen receptor 1 (ESR1) regulon in the 74 BRCA donors profiled in this study identified 482 ATAC-seq distal peak-to-gene links corresponding to 124 ESR1 target genes (fig. S7C and data S8). Accessibility at these peaks is strongly concordant with expectations, further supporting the predicted links ($P < 1 \times 10^{-20}$, fig. S7D). Examination of this regulon across all TCGA

BRCA donors ($N = 1082$) showed a significant difference in overall survival between ESR1 dES-positive and -negative samples (fig. S7, E and F).

Together, pan-cancer and BRCA-specific peak-to-gene links further informed cancer-related GWAS polymorphisms, allowing the linkage of SNPs to putative gene targets with about 65% of all GWAS polymorphisms targeting a gene other than the closest gene on the linear genome (data S5). SNPs falling within peak-to-gene links were predicted to act on important cancer-related genes, including master regulators of cancer and tissue identity such as *NKX2-1* (fig. S7G) and *TP63* (fig. S7H). Focusing specifically on the BRCA peak-to-gene links for which published 3D chromosome conformation data are available, we found clear examples of GWAS SNPs interacting with distant, non-neighboring genes, such as *OSRI* (Fig. 6E and fig. S7I). More-

over, overlapping of the pan-cancer and breast cancer-specific peak-to-gene links with expression quantitative trait loci (eQTLs, where genetic variation at noncoding elements is associated with gene expression differences) from the Genotype-Tissue Expression (GTEx) project showed significant overlap in almost all comparisons ($N = 44$ of 48 comparisons) (fig. S7J and data S5). These results underscored our ability to use these predicted peak-to-gene links to generate key insights into published data and inform poorly understood aspects of cancer biology.

Identification of DNA regulatory elements related to immunological response to cancer

Of particular interest to current cancer therapy, immune infiltrates represent a substantial

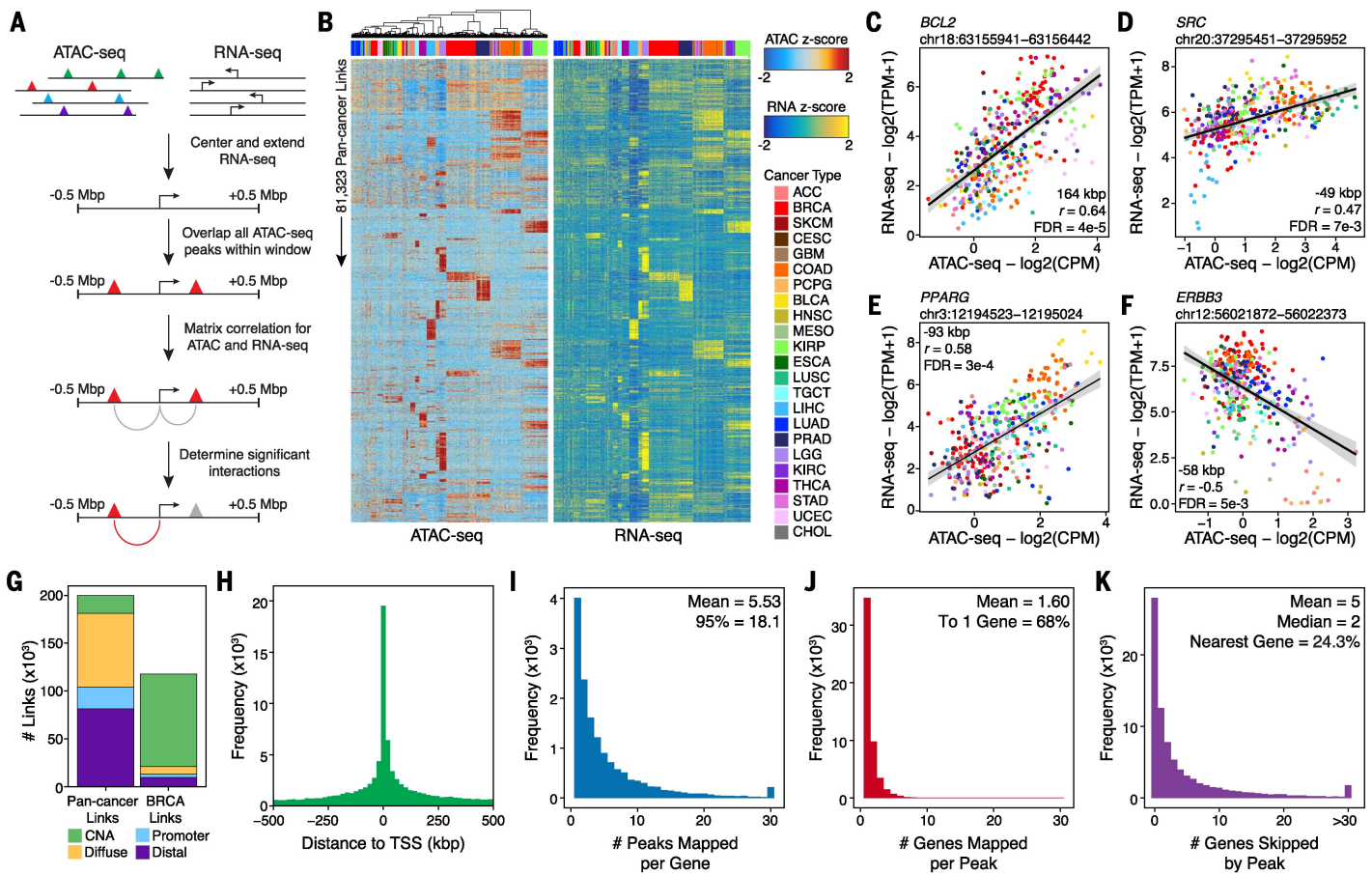


Fig. 5. In silico linking of ATAC-seq peaks to genes. (A) Schematic of the in silico approach used to link ATAC-seq peaks in distal noncoding DNA elements to genes via correlation of chromatin accessibility and RNA expression. (B) Heatmap representation of the 81,323 pan-cancer peak-to-gene links predicted. Each row represents an individual link between one ATAC-seq peak and one gene. Color represents the relative ATAC-seq accessibility (left) or RNA-seq gene expression (right) for each link as a z-score. (C) Dot plot of the ATAC-seq accessibility and RNA-seq gene expression of a peak-to-gene link located 164 kbp away from the transcription start site of the *BCL2* gene (peak 498895) that is predicted to regulate its expression. Color represents the cancer type. Each dot represents an individual sample. (D) Same as in Fig. 5C but for a peak that is located 49 kbp away from the *SRC* gene (peak 525295). (E) Same

as in Fig. 5C but for a peak that is located 93 kbp away from the *PPARG* gene (peak 98874). (F) Same as in Fig. 5C but for a peak that is located 58 kbp away from the *ERBB3* gene (peak 381116). (G) Bar plot showing the number of predicted links that were filtered for various reasons. First, regions whose correlation is driven by DNA copy number amplification were excluded (“CNA”). Next, regions of high local correlation were filtered out (“Diffuse”). Lastly, peak-to-gene links where the peak overlapped a promoter region were excluded (“Promoter”). The remaining links (“Distal”) are used in downstream analyses. (H) Distribution of the distance of each peak to the transcription start site (TSS) of the linked gene. (I) Distribution of the number of peaks linked per gene. (J) Distribution of the number of genes linked per peak. (K) Distribution of the number of genes “skipped” by a peak to reach its predicted linked gene.

contribution to the overall tumor composition in solid tumors (46–48). We reasoned that infiltrating immune cells could contribute to our ATAC-seq data, both through actions on tumor cells and through increased chromatin accessibility at known immune-specific regulatory elements. Leveraging published ATAC-seq datasets from the human hematopoietic system (25) and data generated here from human dendritic cell subsets (Fig. 6F), we characterized each of our linked peaks by comparing its accessibility in immune cell types to its accessibility in bulk cancer samples (Fig. 6G). We reasoned that peaks that are more accessible in immune cells compared with our cancer cohort might be generated from immune cells associated with the tumor tissue (Fig. 6G). Additionally, we correlated each linked peak to the cytolytic activity score (49) of the tumor. The cytolytic activity score is based on the log-average gene expression of granzyme A and perforin 1, two CD8 T cell-specific markers. Linked peaks that exhibit high correlation to cytolytic activity might also be considered to be related to immune infiltration. Combining these two metrics, we identified peak-to-gene links expected to be highly relevant to immune infiltration, including links to genes relevant to antigen presentation and T cell response (Fig. 6H and data S9). The accessibility of these peak-to-gene links that were predicted to be immune-related is highly cor-

related with computationally predicted metrics of immune infiltration (46, 47) and inversely correlated with tumor purity (48) (Fig. 6I). One notable linked gene is programmed death ligand 1 (*PDL1*, also known as *CD274*), a key mediator of immune evasion by cancer and an important target for cancer immunotherapy. *PDL1* is linked to four putative distal regulatory elements that exhibit distinct chromatin accessibility across cancer types and are located as far as 43 kbp away from the *PDL1* transcription start site (Fig. 6, J and K). CRISPRi of each of these four putative *PDL1* regulatory elements significantly decreased, but did not abrogate, the expression of *PDL1* mRNA in at least one of the two breast cancer cell lines tested (MCF7 and MDA-MB-231 cells, Fig. 6L). These results support a model where the expression of *PDL1* is affected by the combined activity of multiple distal regulatory elements.

Identification of cancer-relevant noncoding mutations

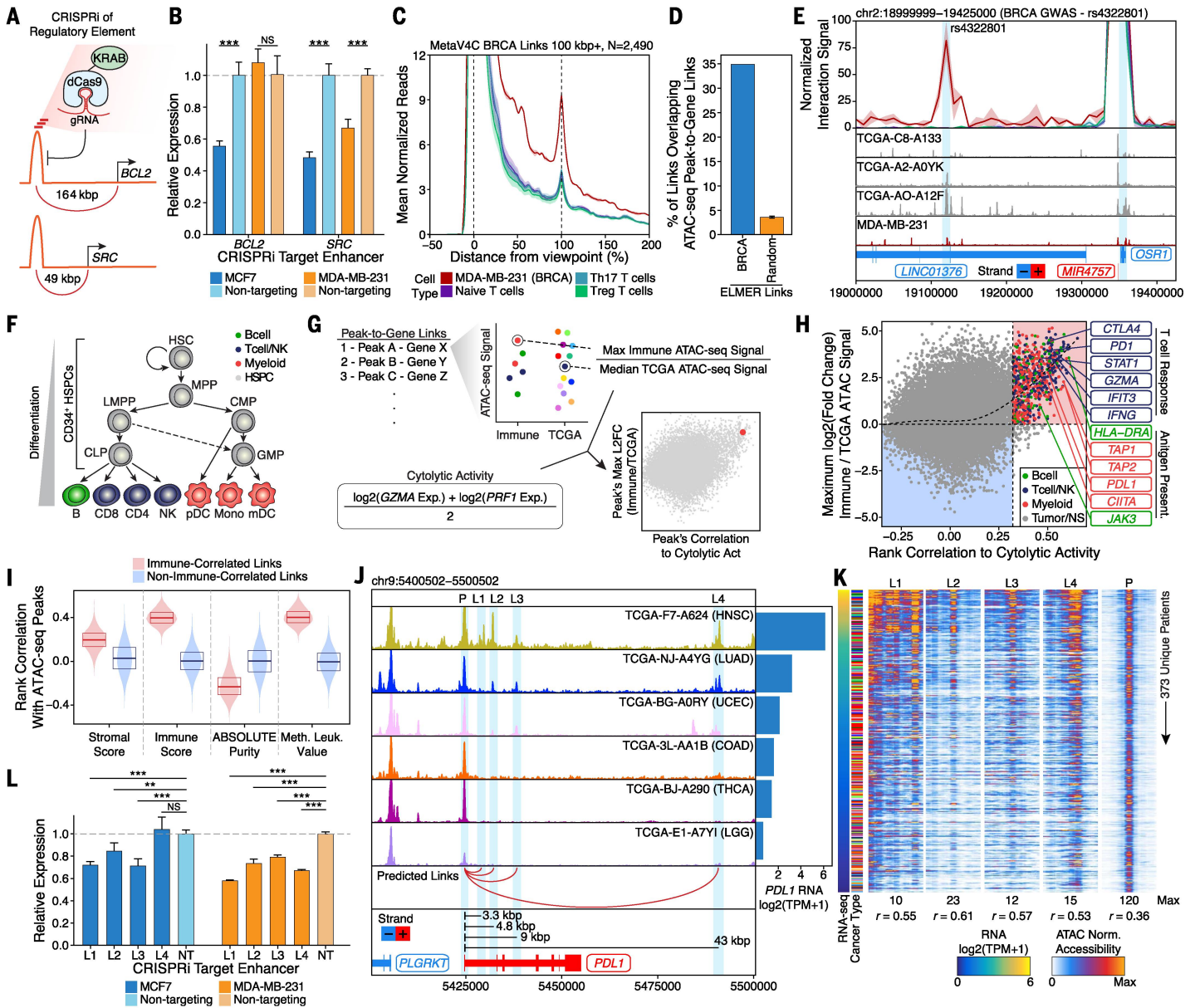
In addition to identifying gene regulatory interactions in cancer, ATAC-seq combined with whole-genome sequencing (WGS) can be used to identify regulatory mutations driving cancer initiation and progression. For example, if a noncoding somatic mutation causes the generation of a TF binding site, this mutation could lead to an increase in chromatin accessibility

in cis and a concomitant increase in the observed frequency of the mutant allele in ATAC-seq as compared with that in WGS (Fig. 7A). Similarly, a mutation that inactivates a TF binding site can lead to a decrease in chromatin accessibility and a concomitant decrease in the observed frequency of the mutant allele. If such mutations in regulatory elements were to be functional in cancer, we might also expect that they increase or decrease chromatin accessibility beyond the expected distribution observed in nonmutated samples.

From the 404 donors profiled in this study, high-depth WGS data was available for 35 donors across 10 cancer types. These 35 donors had 374,705 called somatic mutations, with 32,696 falling within annotated ATAC-seq peaks and 2259 having at least 30 reads in both ATAC-seq and WGS data (data S10). Among these mutations were three separate occurrences of telomerase reverse transcriptase (*TERT*) gene promoter mutations (Fig. 7B), previously shown to generate de novo E26 transformation-specific (ETS) motif sites. ATAC-seq is especially well suited to identifying these *TERT* promoter mutations because the variant allele frequency is skewed owing to the increase in accessibility on the mutant allele (fig. S8A). Compared with the publicly available exome sequencing data from TCGA, where the *TERT* capture probes do not extend into the promoter region, ATAC-seq provided significantly

Fig. 6. Validation of long-range gene regulation of cancer in peak-to-gene links. (A) Schematic of CRISPRi experiments performed. Each experiment uses three guide RNAs (gRNAs) to target an individual peak. The effect of this perturbation on the expression of the linked gene is determined with quantitative polymerase chain reaction (qPCR). (B) Gene expression changes by qPCR after CRISPRi of peaks predicted to be linked to the *BCL2* (peak 498895) and *SRC* (peak 525295) genes in MCF7 and MDA-MB-231 cells. Error bars represent the standard deviation of four technical replicates. *** $P < 0.001$ and NS is not significant by two-tailed Student's *t* test. (C) Meta-virtual circular chromosome conformation capture (4C) plot of predicted BRCA-specific peak-to-gene links with distances greater than 100 kbp. HiChIP interaction frequency is shown for the MDA-MB-231 basal breast cancer cell line as well as multiple populations of primary T cells. Th17, T helper 17 cell; Treg, regulatory T cell. (D) Bar plot showing the overlap of predicted ATAC-seq-based peak-to-gene links and DNA methylation-based ELMER predicted probe-to-gene links in BRCA, as a percentage of all ATAC-seq-based peak-to-gene links with a peak overlapping a methylation probe. The percentage of peak-to-gene links overlapping an ELMER probe-to-gene link (34.9%) is compared to the overlap with 1000 sets of randomized ELMER probe-to-gene links ($3.6 \pm 0.6\%$, $P < 0.001$). (E) Virtual 4C plot of the peak-to-gene link between rs4322801 and the *OSRI* gene. Normalized HiChIP interaction signal is shown for the MDA-MB-231 basal breast cancer cell line as well as multiple populations of primary T cells using the colors shown in Fig. 6C. ATAC-seq sequencing tracks are shown below for four BRCA samples and MDA-MB-231 cells with increasing levels of *OSRI* gene expression. The rs4322801 SNP (left) and *OSRI* gene (right) are highlighted by light blue shading. Region shown represents chr2:18999999 to 19425000. (F) Diagram of the hematopoietic differentiation hierarchy with differentiated cells colored as either B cells (green), T cell or natural killer (Tcell/NK) cells (blue), or myeloid cells (red). HSC, hematopoietic stem cell; LMPP, lymphoid-primed multipotent progenitor; CLP, common lymphoid progenitor; MPP, multipotent progenitor; CMP, common myeloid progenitor; GMP, granulocyte macrophage progenitor; HSPC, hematopoietic stem and progenitor cells; pDC, plasmacy-

toid dendritic cell; mDC, myeloid dendritic cell. (G) Schematic of the analysis shown in Fig. 6H. Peak-to-gene links are classified as related to immune infiltration if their accessibility is higher in immune cells than TCGA cancer samples and they are highly correlated to cytolytic activity. (H) Dot plot showing ATAC-seq peak-to-gene links with relevance to immune infiltration. Each dot represents an individual peak with a predicted gene link. Peaks that are related to immune cells have higher ATAC-seq accessibility in immune cell types compared to TCGA cancer samples. Peaks related to immune infiltration have a higher correlation to cytolytic activity. Color represents the cell type of the observation. The vertical dotted line represents the mean + 2.5 standard deviations above the mean for all ATAC-seq peak correlations to the cytolytic activity. The red shading indicates peak-to-gene links that are predicted to be related to immune infiltration. The blue shading indicates peak-to-gene links that are not predicted to be related to immune infiltration. NS, not significant. (I) Violin plots of the distribution of Spearman correlations across all peak-to-gene links predicted to be related to immune infiltration (red) or not (blue) with various metrics of tumor purity. (J) Normalized ATAC-seq sequencing tracks of the *PDL1* gene locus in six samples with variable levels of expression of the *PDL1* gene (right). Predicted links (red) are shown below for four peak-to-gene links (L1 to L4, peaks 293734, 293735, 293736, and 293740, respectively) to the promoter of *PDL1*. One of these peak-to-gene links (L2) overlaps an alternative start site for *PDL1* and was therefore labeled as a “promoter” peak during filtration. This peak-to-gene link was added to this analysis after manual observation. Region shown represents chr9:5400502 to 5500502. (K) Heatmap representation of the ATAC-seq chromatin accessibility of the 5000-bp region centered at each of the four peak-to-gene links shown in Fig. 6J. Each row represents a unique donor ($N = 373$) ranked by *PDL1* expression. The correlation of the chromatin accessibility of each peak with the expression of *PDL1* is shown below the plot. Color represents normalized accessibility. (L) Gene expression changes of the *PDL1* gene by qPCR after CRISPRi of peaks predicted to be linked to the *PDL1* gene in MCF7 and MDA-MB-231 cells. Error bars represent the standard deviation of four technical replicates. *** $P < 0.0001$ and ** $P < 0.05$ by two-tailed Student's *t* test.



higher sequencing coverage of the *TERT* promoter locus per read sequenced, enabling a more robust classification of *TERT* promoter mutations ($P < 1 \times 10^{-7}$, fig. S8B). Of the three *TERT* promoter mutations identified in the subset of donors with matched WGS, one mutation, in particular, leads to a significant increase in accessibility compared to the other nonmutated members of that cancer type (FDR < 0.0001, blue dot in Fig. 7, B and C). As expected, this increase in *TERT* promoter accessibility is associated with a concomitant increase in *TERT* gene expression (blue dot in Fig. 7C). *TERT* promoter mutations, however, are not the only way to increase *TERT* gene expression, because high *TERT* expression can also be observed in samples without identifiable *TERT* promoter mutations (Fig. 7C). Consistent with a previous report (50), differential motif analysis at the site of this *TERT* promoter mutation identified E74-like ETS tran-

scription factor 1 (ELF1) or ELF2 as the TF that likely binds to the de novo ETS motif (fig. S8C). In addition, we identified several mutations overlapping CCCTC-binding factor (CTCF) motif occurrences that are associated with decreased accessibility at that site (fig. S8, D and E). However, these mutations were relatively rare and often had only small effects on the accessibility of the CTCF motif site despite a known enrichment of somatic mutations in CTCF motif sites in cancer (51, 52).

In addition to known *TERT* promoter mutations, integrative analysis of WGS and ATAC-seq data uncovered a mutation upstream of the FYVE RhoGEF and PH domain-containing 4 (*FGD4*) gene, a regulator of the actin cytoskeleton and cell shape. This mutation occurs in a bladder cancer sample where the variant allele frequency observed in ATAC-seq is markedly higher than the variant allele frequency observed in

WGS (Fig. 7B). This mutation is associated with a significant increase in accessibility compared to other bladder cancer samples in this cohort (Fig. 7, B and D) and is accompanied by a similar increase in *FGD4* mRNA (Fig. 7D). Moreover, this mutation upstream of the *FGD4* gene (referred to as eFGD4 for enhancer FGD4) leads to a level of accessibility that is higher than any of the other samples profiled by ATAC-seq in this study (fig. S8F) and a level of *FGD4* gene expression that is in the top 3% of all bladder cancer samples in TCGA (fig. S8G). As estimated by WGS data, this eFGD4 mutation is present in a subclone comprising about 13% of the tumor (Fig. 7E); however, the mutant allele is present in 96% of all ATAC-seq reads spanning this locus (Fig. 7E), demonstrating a strong preference for accessibility on the mutant allele. This eFGD4 mutation is analogous to, but potentially more potent than, the *TERT* promoter mutation

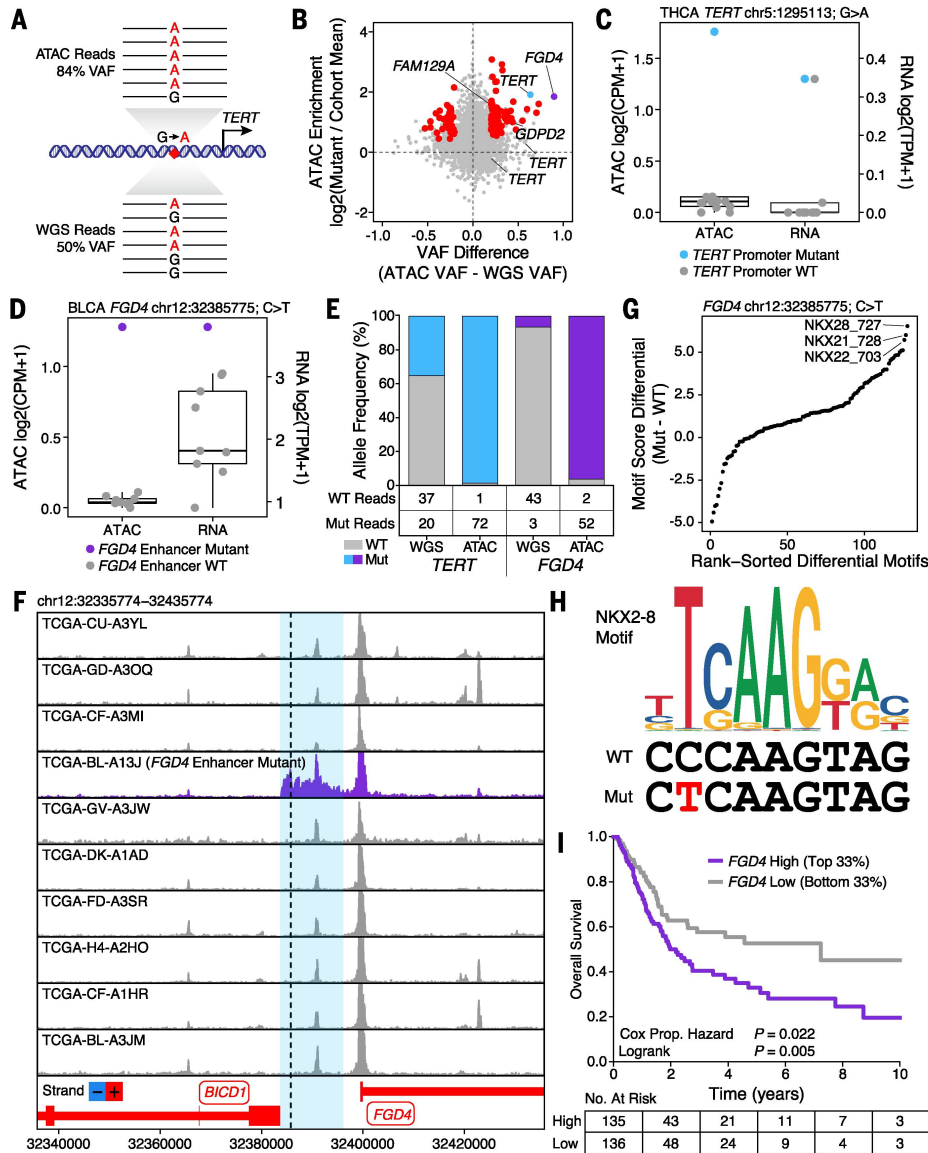


Fig. 7. Integration of WGS and ATAC-seq identifies cancer-relevant regulatory mutations.

(A) Schematic of how functional variants are identified in regulatory elements. The example shown depicts the *TERT* promoter. (B) Dot plot of the difference in variant allele frequency (VAF) of ATAC-seq and WGS and the changes in chromatin accessibility caused by the given variant with respect to other samples of the same cancer type. Variants with a higher variant allele frequency in ATAC-seq than WGS would be expected to cause an increase in accessibility. Each dot represents an individual somatic mutation. (C) Normalized ATAC-seq and RNA-seq of thyroid cancer donors profiled in this study. Each dot represents an individual donor. Blue dot represents the donor with a *TERT* promoter mutation shown in Fig. 7B. Other thyroid cancer donors known to harbor a *TERT* promoter mutation were excluded from this plot. The hinges of the box represent the 25th to 75th percentile. WT, wild type. (D) Normalized ATAC-seq and RNA-seq of bladder cancer donors profiled in this study. Each dot represents an individual donor. Purple dot represents the donor with a mutation upstream of the *FGD4* gene shown in Fig. 7B. The hinges of the box represent the 25th to 75th percentile. (E) Comparison of wild-type and mutant reads in WGS and ATAC-seq data at the *TERT* promoter and *FGD4* upstream region for the donors highlighted in (D) and (E). (F) Normalized ATAC-seq sequencing tracks of the *FGD4* locus in the 10 bladder cancer samples profiled in this study, including the one sample with a mutation predicted to generate a de novo NKX motif (TCGA-BL-A13J). Locus shown represents chr12:32335774 to 32435774. The mutation position is indicated by a black dotted line. The predicted enhancer region surrounding this mutation is highlighted by light blue shading. (G) Difference in motif score in the wild-type and mutant *FGD4* upstream region. Motif score represents the degree of similarity between the sequence of interest and the relevant motif. Each dot represents an individual motif. (H) Overlay of the NKX2-8 motif (CIS-BP M6377_1.02) and the wild-type and mutant sequences of the *FGD4* upstream region. (I) Kaplan-Meier survival analysis of TCGA bladder cancer patients with high (top 33%) and low (bottom 33%) expression for the *FGD4* gene.

described above (Fig. 7E). In the case of the e*FGD4* mutation, this dramatic allele bias occurs because chromatin at this locus is not normally accessible in any of the bladder cancer samples profiled in this study (gray dots and tracks in Fig. 7, D and F) but becomes highly accessible in the context of the e*FGD4* mutation (purple dot and track in Fig. 7, D and F). Differential motif analysis identified NKX factor motifs as the most strongly enriched in the sequence corresponding to the e*FGD4* mutation (Fig. 7G), where a C-to-T transition at position two generated a perfect NKX2-8 motif de novo from a latent site (Fig. 7H). RNA-seq data from the mutated sample identified multiple expressed NKX TFs [transcripts per million (TPM) > 0.5], nominating NKX3-1, NKX2-3, and NKX2-5 as potential mediators of this DNA binding event (fig. S8H). From this, we hypothesized that the e*FGD4* mutation creates a de novo binding site for an NKX TF which, upon binding to the DNA, leads to a broad increase in accessibility across the entire 12-kbp region upstream of the *FGD4* gene. This hypothesis was further supported by the observation that the ATAC-seq accessibility of the entire *FGD4* upstream locus occurs on a single phased allele (fig. S8I). Moreover, separation of subnucleosomal and nucleosome-spanning reads in the ATAC-seq data are consistent with protein binding at the site of the e*FGD4* mutation (light blue shading in fig. S8I). Lastly, because higher *FGD4* expression is significantly associated with worse overall survival in bladder cancer (Fig. 7I and fig. S8J), this mutation could have functional consequence in this particular cancer. Whether the e*FGD4* mutation or other enhancer mutations emerge as recurrent drivers of human cancer should be addressed in future studies. Our data identified multiple additional noncoding mutations associated with a concomitant gain of chromatin accessibility and increase in RNA expression (fig. S8, K to Q), and we anticipate that future work will uncover mechanisms underlying this type of regulatory mutation across all cancer types.

Discussion

Here we provide an initial characterization of the chromatin regulatory landscape in primary human cancers. This dataset identified hundreds of thousands of accessible DNA elements, expanding the dictionary of regulatory elements discovered through previous large-scale efforts such as The Roadmap Epigenomics Project. The identification of these additional elements was made possible through (i) our analysis of primary cancer specimens, (ii) greater saturation of some cancer and tissue types in our dataset, or (iii) potential differences between ATAC-seq and DNase-seq platforms. Nevertheless, the high overlap between the two datasets demonstrates the robustness of both platforms and the consistency of the observed results.

The exquisite cell type-specificity of distal regulatory elements from our ATAC-seq data enabled the classification of cancer types and the discovery of previously unappreciated cancer

subtypes. De novo clustering of TCGA samples based on chromatin accessibility strongly overlaps previous integrative clustering methods, identifying 18 distinct cancer clusters. Comparing this clustering scheme to other clustering schemes defined by cancer type, mRNA, miRNA, DNA methylation, RPPA, and DNA copy number alterations, we observed the strongest concordance of our clustering scheme with mRNA and cancer type, consistent with a close functional linkage between chromatin accessibility and transcriptional output. The strength of the observed associations is influenced by the features represented for each platform. For example, the DNA methylation clusters are based on cancer-specific promoter hypermethylation (28). Clustering based on DNA methylation at distal regulatory elements would likely show a stronger correlation with the ATAC-seq groupings, but distal regulatory element representation on the DNA methylation array used for these samples was too sparse to allow such an analysis. We also identified epigenetically distinct subtypes of kidney renal papillary cancer that have clear differences in overall survival. This cancer type-specific activity in DNA regulatory elements may arise via mutations within the regulatory element, pathologic transcription factor activity, or reflect the regulatory state of the tumor's cell of origin (e.g., stem cells). As the chromatin accessibility landscapes of additional primary cancer samples are profiled, we anticipate the identification of further epigenetic subdivisions with prognostic implications, potentially nominating avenues for therapeutic intervention.

The data generated in this study fully represents the cellular complexity of primary human tumors, comprising signals from tumor cells, infiltrating immune cells, stromal cells, and other normal cell types. In many ways, this complexity is advantageous because it allows complex systems-level analyses to be performed in the future, including cellular deconvolution approaches to understand the contributions of various cell types or cell states to the overall landscape of chromatin accessibility. However, the admixed nature of this signal also highlights the need for future work to profile the chromatin accessibility of matched healthy tissues to further refine the specific changes that drive cancer. Nevertheless, the chromatin accessibility profiles generated in this study represent the largest effort to date to characterize the regulatory landscape in primary human cancer cells.

Using this data-rich resource, we identified classes of TFs whose expression leads to different patterns in TF occupancy and motif protection. By integrating RNA-seq and ATAC-seq, we found factors whose expression is sufficient for both motif protection and nucleosome repositioning and demonstrated this binding to be inversely correlated with the level of DNA methylation at those binding sites. Despite this strong correlation, many sites of differential chromatin accessibility do not show differential methylation, demonstrating the complemen-

tarity of these two data types, perhaps owing to the presence of intermediate chromatin states such as poised promoters or enhancers (53, 54).

Moreover, integration of RNA-seq and ATAC-seq across the 373 donors with paired datasets enabled a quantitative model to link the accessibility of a regulatory element to the expression of predicted target genes. This workflow identified putative links for more than half of the protein-coding genes in the genome, informing the target genes of poorly understood GWAS SNPs and increasing our understanding of cancer gene regulatory networks. These predictions were further supported using 3D chromosome conformation data, and a subset were validated through CRISPRi experiments in breast cancer cell lines. However, profiling of chromosome conformation in primary cancer samples has not been performed on a large scale. Future work to produce maps of chromosome conformation in these or other primary cancer samples will improve our understanding of gene regulatory networks in cancer and further clarify the roles for certain GWAS-identified SNPs in cancer initiation and progression.

Lastly, through integration of WGS and ATAC-seq, we revealed a class of somatic mutations that occur in regulatory regions and lead to strong gains in chromatin accessibility. We demonstrated that these mutations likely lead to changes in nearby gene expression and affect genes whose expression is linked to poorer overall survival. Some of these mutations, such as those occurring in the *TERT* promoter, have been found to be recurrent whereas others, such as the mutation upstream of the *FGD4* gene, may be rare but functionally important. Because the enhancer functions are often distributed and latent enhancer sequences are pervasive in the genome, noncoding mutations in cancer may be especially challenging and require high-throughput functional assessment. Future larger-scale efforts to combine genome and epigenome sequencing will pave the way to tackling the noncoding genome in cancer.

Materials and methods summary

ATAC-seq data was generated from 410 tissue samples from the TCGA collection of primary human tumors. These samples spanned 23 different tumor types. These ATAC-seq data were used to cluster samples, identifying epigenetically defined patient subgroups. Moreover, TF regulators of cancer were defined, and footprinting of these regulators was correlated to gene expression to identify putative classes of TFs. A correlation-based model was developed to link ATAC-seq peaks to putative target genes. These putative links were validated using CRISPRi-based perturbation of the peak region followed by quantification of changes in gene expression. Publicly available HiChIP data and GTEx eQTL data were further used to support genome-wide peak-to-gene linkage predictions. Lastly, WGS and ATAC-seq were combined to identify noncoding mutations that affect chromatin accessibility in an allele-specific manner.

REFERENCES AND NOTES

- C. Hutter, J. C. Zenklusen, The Cancer Genome Atlas: Creating lasting value beyond its data. *Cell* **173**, 283–285 (2018). doi: [10.1016/j.cell.2018.03.042](https://doi.org/10.1016/j.cell.2018.03.042); pmid: [29625045](https://pubmed.ncbi.nlm.nih.gov/29625045/)
- W. A. Flavahan, E. Gaskell, B. E. Bernstein, Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, eaal2380 (2017). doi: [10.1126/science.aal2380](https://doi.org/10.1126/science.aal2380); pmid: [28729483](https://pubmed.ncbi.nlm.nih.gov/28729483/)
- D. Hanahan, R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011). doi: [10.1016/j.cell.2011.02.013](https://doi.org/10.1016/j.cell.2011.02.013); pmid: [21376230](https://pubmed.ncbi.nlm.nih.gov/21376230/)
- M. Egeblad, E. S. Nakasone, Z. Werb, Tumors as organs: Complex tissues that interface with the entire organism. *Dev. Cell* **18**, 884–901 (2010). doi: [10.1016/j.devcel.2010.05.012](https://doi.org/10.1016/j.devcel.2010.05.012); pmid: [20627072](https://pubmed.ncbi.nlm.nih.gov/20627072/)
- W. Zhou et al., DNA methylation loss in late-replicating domains is linked to mitotic cell division. *Nat. Genet.* **50**, 591–602 (2018). doi: [10.1038/s41588-018-0073-4](https://doi.org/10.1038/s41588-018-0073-4); pmid: [29610480](https://pubmed.ncbi.nlm.nih.gov/29610480/)
- M. Almamun et al., Integrated methylome and transcriptome analysis reveals novel regulatory elements in pediatric acute lymphoblastic leukemia. *Epigenetics* **10**, 882–890 (2015). doi: [10.1080/15592294.2015.1078050](https://doi.org/10.1080/15592294.2015.1078050); pmid: [26308964](https://pubmed.ncbi.nlm.nih.gov/26308964/)
- Y. He et al., Improved regulatory element prediction based on tissue-specific local epigenomic signatures. *Proc. Natl. Acad. Sci. U.S.A.* **114**, E1633–E1640 (2017). doi: [10.1073/pnas.1618353114](https://doi.org/10.1073/pnas.1618353114); pmid: [28193886](https://pubmed.ncbi.nlm.nih.gov/28193886/)
- L. Yao, H. Shen, P. W. Laird, P. J. Farnham, B. P. Berman, Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. *Genome Biol.* **16**, 105 (2015). doi: [10.1186/s13059-015-0668-3](https://doi.org/10.1186/s13059-015-0668-3); pmid: [25994056](https://pubmed.ncbi.nlm.nih.gov/25994056/)
- M. Ceccarelli et al., Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. *Cell* **164**, 550–563 (2016). doi: [10.1016/j.cell.2015.12.028](https://doi.org/10.1016/j.cell.2015.12.028); pmid: [26824661](https://pubmed.ncbi.nlm.nih.gov/26824661/)
- H. Noshmeh et al., Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell* **17**, 510–522 (2010). doi: [10.1016/j.ccr.2010.03.017](https://doi.org/10.1016/j.ccr.2010.03.017); pmid: [20399149](https://pubmed.ncbi.nlm.nih.gov/20399149/)
- T. Hinoue et al., Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22**, 271–282 (2012). doi: [10.1101/gr.117523.110](https://doi.org/10.1101/gr.117523.110); pmid: [21659424](https://pubmed.ncbi.nlm.nih.gov/21659424/)
- P. A. Northcott et al., The whole-genome landscape of medulloblastoma subtypes. *Nature* **547**, 311–317 (2017). doi: [10.1038/nature22973](https://doi.org/10.1038/nature22973); pmid: [28726821](https://pubmed.ncbi.nlm.nih.gov/28726821/)
- The Cancer Genome Atlas Research Network, Comprehensive molecular characterization of papillary renal-cell carcinoma. *N. Engl. J. Med.* **374**, 135–145 (2016). doi: [10.1056/NEJMoa1505917](https://doi.org/10.1056/NEJMoa1505917); pmid: [26536169](https://pubmed.ncbi.nlm.nih.gov/26536169/)
- B. Akhtar-Zaidi et al., Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736–739 (2012). doi: [10.1126/science.1217277](https://doi.org/10.1126/science.1217277); pmid: [22499810](https://pubmed.ncbi.nlm.nih.gov/22499810/)
- H. Chen et al., A pan-cancer analysis of enhancer expression in nearly 9000 patient samples. *Cell* **173**, 386–399.e12 (2018). doi: [10.1016/j.cell.2018.03.027](https://doi.org/10.1016/j.cell.2018.03.027); pmid: [29625054](https://pubmed.ncbi.nlm.nih.gov/29625054/)
- J. D. Buenostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, W. J. Greenleaf, Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013). doi: [10.1038/nmeth.2688](https://doi.org/10.1038/nmeth.2688); pmid: [24097267](https://pubmed.ncbi.nlm.nih.gov/24097267/)
- M. R. Corces et al., An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017). doi: [10.1038/nmeth.4396](https://doi.org/10.1038/nmeth.4396); pmid: [28846090](https://pubmed.ncbi.nlm.nih.gov/28846090/)
- A. Kundaje et al., Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015). doi: [10.1038/nature14248](https://doi.org/10.1038/nature14248); pmid: [25693563](https://pubmed.ncbi.nlm.nih.gov/25693563/)
- J. Schuijers et al., Transcriptional dysregulation of MYC reveals common enhancer-docking mechanism. *Cell Reports* **23**, 349–360 (2018). doi: [10.1016/j.celrep.2018.03.056](https://doi.org/10.1016/j.celrep.2018.03.056); pmid: [29641996](https://pubmed.ncbi.nlm.nih.gov/29641996/)
- G. Andrey et al., A switch between topological domains underlies *HoxD* genes collinearity in mouse limbs. *Science* **340**, 1234167 (2013). doi: [10.1126/science.1234167](https://doi.org/10.1126/science.1234167); pmid: [23744951](https://pubmed.ncbi.nlm.nih.gov/23744951/)
- M. Yeager et al., Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007). doi: [10.1038/ng2022](https://doi.org/10.1038/ng2022); pmid: [17401363](https://pubmed.ncbi.nlm.nih.gov/17401363/)
- I. P. M. Tomlinson et al., A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008). doi: [10.1038/ng.111](https://doi.org/10.1038/ng.111); pmid: [18372905](https://pubmed.ncbi.nlm.nih.gov/18372905/)

23. I. K. Sur *et al.*, Mice lacking a *Myc* enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360–1363 (2012). doi: [10.1126/science.1228606](https://doi.org/10.1126/science.1228606); pmid: 23118011
24. R. E. Thurman *et al.*, The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012). doi: [10.1038/nature11232](https://doi.org/10.1038/nature11232); pmid: 22955617
25. M. R. Corces *et al.*, Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016). doi: [10.1038/ng.3646](https://doi.org/10.1038/ng.3646); pmid: 27526324
26. L. J. P. van der Maaten, G. E. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
27. A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014). doi: [10.1126/science.1242072](https://doi.org/10.1126/science.1242072); pmid: 24970081
28. K. A. Hoadley *et al.*, Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* **173**, 291–304.e6 (2018). doi: [10.1016/j.cell.2018.03.022](https://doi.org/10.1016/j.cell.2018.03.022); pmid: 29625048
29. M. Kulis *et al.*, Whole-genome fingerprint of the DNA methylome during human B cell differentiation. *Nat. Genet.* **47**, 746–756 (2015). doi: [10.1038/ng.3291](https://doi.org/10.1038/ng.3291); pmid: 26053498
30. H. S. Kim *et al.*, Pluripotency factors functionally premark cell-type-restricted enhancers in ES cells. *Nature* **556**, 510–514 (2018). doi: [10.1038/s41586-018-0048-8](https://doi.org/10.1038/s41586-018-0048-8); pmid: 29670286
31. H. Shen *et al.*, Integrated molecular characterization of testicular germ cell tumors. *Cell Reports* **23**, 3392–3406 (2018). doi: [10.1016/j.celrep.2018.05.039](https://doi.org/10.1016/j.celrep.2018.05.039); pmid: 29898407
32. S. Werner *et al.*, Dual roles of the transcription factor grainyhead-like 2 (GRHL2) in breast cancer. *J. Biol. Chem.* **288**, 22993–23008 (2013). doi: [10.1074/jbc.M113.456293](https://doi.org/10.1074/jbc.M113.456293); pmid: 23814079
33. S. K. Denny *et al.*, Nf1b promotes metastasis through a widespread increase in chromatin accessibility. *Cell* **166**, 328–342 (2016). doi: [10.1016/j.cell.2016.05.052](https://doi.org/10.1016/j.cell.2016.05.052); pmid: 27374332
34. S. Baek, I. Goldstein, G. L. Hager, Bivariate genomic footprinting detects changes in transcription factor activity. *Cell Reports* **19**, 1710–1722 (2017). doi: [10.1016/j.celrep.2017.05.003](https://doi.org/10.1016/j.celrep.2017.05.003); pmid: 28538187
35. A. N. Schep, B. Wu, J. D. Buenrostro, W. J. Greenleaf, ChromVAR: Inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017). doi: [10.1038/nmeth.4401](https://doi.org/10.1038/nmeth.4401); pmid: 28825706
36. Y. Yin *et al.*, Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, eaaj2239 (2017). doi: [10.1126/science.aaj2239](https://doi.org/10.1126/science.aaj2239); pmid: 28473536
37. T. Ellis *et al.*, The transcriptional repressor CDP (Cut11) is essential for epithelial cell differentiation of the lung and the hair follicle. *Genes Dev.* **15**, 2307–2319 (2001). doi: [10.1101/gad.200101](https://doi.org/10.1101/gad.200101); pmid: 11544187
38. B. M. Javierre *et al.*, Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384.e19 (2016). doi: [10.1016/j.cell.2016.09.037](https://doi.org/10.1016/j.cell.2016.09.037); pmid: 27863249
39. C. P. Fulco *et al.*, Systematic mapping of functional enhancer-promoter connections with CRISPR interference. *Science* **354**, 769–773 (2016). doi: [10.1126/science.aag2445](https://doi.org/10.1126/science.aag2445); pmid: 27708057
40. L. S. Qi *et al.*, Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013). doi: [10.1016/j.cell.2013.02.022](https://doi.org/10.1016/j.cell.2013.02.022); pmid: 23452860
41. Y. H. Eom, H. S. Kim, A. Lee, B. J. Song, B. J. Chae, BCL2 as a subtype-specific prognostic marker for breast cancer. *J. Breast Cancer* **19**, 252–260 (2016). doi: [10.4048/jbc.2016.19.3.252](https://doi.org/10.4048/jbc.2016.19.3.252); pmid: 27721874
42. S. W. Cho *et al.*, Promoter of lncRNA gene *PVT1* is a tumor-suppressor DNA boundary element. *Cell* **173**, 1398–1412.e22 (2018). doi: [10.1016/j.cell.2018.03.068](https://doi.org/10.1016/j.cell.2018.03.068); pmid: 29731168
43. T. C. Silva, S. G. Coetzee, L. Yao, D. J. Hazelett, H. Noushimehr, B. P. Berman, Enhancer linking by methylation/expression relationships with the R package ELMER version 2. bioRxiv 148726 [Preprint]. 11 June 2017. doi: [10.1101/148726](https://doi.org/10.1101/148726)
44. A. G. Robertson *et al.*, Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell* **171**, 540–556.e25 (2017). doi: [10.1016/j.cell.2017.09.007](https://doi.org/10.1016/j.cell.2017.09.007); pmid: 28988769
45. M. A. A. Castro *et al.*, Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2016). doi: [10.1038/ng.3458](https://doi.org/10.1038/ng.3458); pmid: 26618344
46. V. Thorsson *et al.*, The immune landscape of cancer. *Immunity* **48**, 812–830.e14 (2018). doi: [10.1016/j.immuni.2018.03.023](https://doi.org/10.1016/j.immuni.2018.03.023); pmid: 29628290
47. K. Yoshihara *et al.*, Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013). doi: [10.1038/ncomms3612](https://doi.org/10.1038/ncomms3612); pmid: 24113773
48. S. L. Carter *et al.*, Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012). doi: [10.1038/nbt.2203](https://doi.org/10.1038/nbt.2203); pmid: 22544022
49. M. S. Rooney, S. A. Shukla, C. J. Wu, G. Getz, N. Hacohen, Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell* **160**, 48–61 (2015). doi: [10.1016/j.cell.2014.12.033](https://doi.org/10.1016/j.cell.2014.12.033); pmid: 25594174
50. M. M. Makowski *et al.*, An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics* **16**, 417–426 (2016). doi: [10.1002/pmic.201500327](https://doi.org/10.1002/pmic.201500327); pmid: 26553150
51. R. Katainen *et al.*, CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015). doi: [10.1038/ng.3335](https://doi.org/10.1038/ng.3335); pmid: 26053496
52. D. Hnisz *et al.*, Activation of proto-oncogenes by disruption of chromosome neighborhoods. *Science* **351**, 1454–1458 (2016). doi: [10.1126/science.aad9024](https://doi.org/10.1126/science.aad9024); pmid: 26940867
53. R. D. Hawkins *et al.*, Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6**, 479–491 (2010). doi: [10.1016/j.stem.2010.03.018](https://doi.org/10.1016/j.stem.2010.03.018); pmid: 20452322
54. T. K. Kelly *et al.*, Genome-wide mapping of nucleosome positioning and DNA methylation within individual DNA molecules. *Genome Res.* **22**, 2497–2506 (2012). doi: [10.1101/gr.143008.112](https://doi.org/10.1101/gr.143008.112); pmid: 22960375

ACKNOWLEDGMENTS

We thank X. Ji and J. Collier for assistance in sequencing, P. Giresi and Epinomics for sharing advice and expertise related to ATAC-seq data analysis, and the members of the Greenleaf and Chang laboratories for thoughtful advice and critique.

Funding: Supported by the National Cancer Institute, NIH grants R35-CA209919 (to H.Y.C.), P50-HG007735 (to H.Y.C. and W.J.G.), and the Parker Institute for Cancer Immunotherapy (H.Y.C.). M.R.C. is supported by NIH K99-AG059918. Additional support through the NIH Genomic Data Analysis Networks 1U24CA210974-01 (J. Zhu), 1U24CA210949-01 (J. N. Weinstein), 1U24CA210978-01 (R. Beroukhim), 1U24CA210952-01 (S. J. Jones), 1U24CA210989-01 (O. Elemento), 1U24CA210990-01 (J. Stuart), 1U24CA210950-01 (R. Akbani), 1U24CA210969-01 (P.W.L.), and 1U24CA210988-01 (K.A.H.). W.J.G. is a Chan-Zuckerberg Biohub Investigator. H.Y.C. is an Investigator of the Howard Hughes Medical Institute. **Author contributions:** L.M.S., J.C.Z., W.J.G., and H.Y.C. conceived of and designed the project. M.R.C. and J.M.G. compiled figures and wrote the manuscript with the help of all authors. M.R.C. developed methodology for profiling frozen cancer tissues by ATAC-seq. S.S., B.H.L., and M.R.C. performed all tissue processing and ATAC-seq data generation. N.C.S. designed and wrote the ATAC-seq data processing pipeline with help from M.R.C. M.R.C. and J.M.G. processed all ATAC-seq data, and J.M.G.

performed all analyses and developed all analytical tools unless otherwise stated below. J.A.S. and C.G. performed survival analyses with supervision from C.C., A.G.R., and M.A.C. J.A.S. performed subtyping analysis for KIRP. W.Z. performed WGBS methylation analysis and variation of information clustering analysis with supervision from P.W.L. T.C.S. performed all ELMER analyses with supervision from B.P.B. C.G. performed all regulon analysis with supervision from A.G.R. and M.A.C. C.K.W. performed tumor map analysis. K.A.H. performed cluster coincidence analysis comparing ATAC-seq-derived clusters to TCGA iClusters. S.W.C. produced all Tn5 transposase used in this study and generated reagents and cell lines used in CRISPRi experiments. B.H.L., S.S., and M.R.C. performed CRISPRi experiments. A.T.S. generated human dendritic cell ATAC-seq data. J.M.G. and A.T.S. performed immune infiltration analysis. J.M.G. and M.R.M. performed HiChIP analysis. M.R.C. performed all analysis for identifying noncoding mutations from WGS and ATAC-seq data. I.F. coordinated all TCGA analysis working group efforts. J.C.Z. selected tumor samples to profile in this study. P.W.L. and W.J.G. co-chaired the TCGA analysis working group. C.C. provided expertise relevant to pan-cancer data analysis. H.Y.C. and W.J.G. supervised overall data generation and analysis. All authors listed under “The Cancer Genome Atlas Analysis Network” provided valuable input and expertise. **Competing interests:** H.Y.C. is a co-founder of Accent Therapeutics and Epinomics and is an adviser of 10X Genomics and Spring Discovery. W.J.G. is a co-founder of Epinomics and an adviser to 10X Genomics, Guardant Health, Centrillion, and NuGen. C.C. is an adviser to GRAIL. Stanford University holds a patent on ATAC-seq, on which H.Y.C. and W.J.G. are named as inventors. **Data and materials availability:** Processed data not provided in the supplementary data files are available through our TCGA Publication Page (<https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>). This includes pan-cancer raw and normalized counts matrices, cancer type-specific peak calls, cancer type-specific raw and normalized count matrices, and bigWig track files for all technical replicates. Raw ATAC-seq data as fastq or aligned BAM files will be made available through the NIH Genomic Data Commons portal (<https://portal.gdc.cancer.gov/>). ATAC-seq data corresponding to human plasmacytoid dendritic cells and myeloid dendritic cells (the only non-TCGA data generated here) is available through SRA BioProject PRJNA491478. The ATAC-seq peak accessibility and computed peak-to-gene linkage predictions are publicly available for interactive visualization and exploration at the UCSC Xena Browser (<https://ataseq.xenahubs.net>). Sample-level ATAC-seq data across all 404 donors assayed can be visualized side-by-side with all other data from TCGA, including gene expression, DNA methylation from both Illumina 450K array and WGBS platforms, and ELMER enhancer analysis results, as well as the latest survival data and mutation calls from the Genomic Data Commons. ATAC-seq data can be queried by gene, genomic position, or individual peaks. The UCSC Xena Browser makes this rich resource available for interactive online analysis and visualization by the larger scientific community. Samples from the TCGA project can only be used for TCGA efforts owing to restrictions in the material transfer agreement used for acquisition. No external groups can access the tissue or analytes.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/362/6413/eaav1898/suppl/DC1
TCGA Analysis Network Collaborators
Materials and Methods
Protocol S1
Figs. S1 to S8
References (55–78)
Data S1 to S10

22 August 2018; accepted 28 September 2018
10.1126/science.aav1898

ANEXO II - *RTNDUALS*: AN R/BIOCONDUCTOR PACKAGE FOR ANALYSIS OF CO-REGULATION
AND INFERENCE OF DUAL REGULONS



Systems biology

RTNduals*: An R/Bioconductor package for analysis of co-regulation and inference of *dual regulons

Vinicius S. Chagas^{1,†}, Clarice S. Groeneveld^{1,†}, Kelin G. Oliveira^{1,2}, Sheyla Trefflich³, Rodrigo C. de Almeida⁴, Bruce A. J. Ponder⁵, Kerstin B. Meyer^{5,6}, Steven J. M. Jones⁷, A. Gordon Robertson^{7,‡,*} and Mauro A. A. Castro^{1,‡,*}

¹Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba, 81520-260, Brazil, ²Department of Clinical Sciences, Section of Oncology and Pathology, Lund University, Lund, 221 85, Sweden, ³Bioinformatics Department, Federal University of Minas Gerais, Belo Horizonte, 31270-901, Brazil, ⁴Department of Biomedical Data Sciences, Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands, ⁵Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, United Kingdom, ⁶Wellcome Sanger Institute, Hinxton, CB10 1SA, United Kingdom and ⁷Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, V5Z 4S6, Canada

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

‡The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: XXXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Transcription factors (TFs) are key regulators of gene expression, and can activate or repress multiple target genes, forming regulatory units, or regulons. Understanding downstream effects of these regulators includes evaluating how TFs cooperate or compete within regulatory networks. Here we present *RTNduals*, an R/Bioconductor package that implements a general method for analysing pairs of regulons.

Results: *RTNduals* identifies a dual regulon when the number of targets shared between a pair of regulators is statistically significant. The package extends the *RTN* (Reconstruction of Transcriptional Networks) package, and uses *RTN* transcriptional networks to identify significant co-regulatory associations between regulons. The Supplementary Information reports two case studies for TFs using the METABRIC and TCGA breast cancer cohorts.

Availability: *RTNduals* is written in the R language, and is available from the Bioconductor project at <http://bioconductor.org/packages/RTNduals/>

Contact: mauro.castro@ufpr.br, grobertson@bcgsc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene regulation in eukaryotes integrates a large number of interconnected regulatory influences. Some of the major contributors in gene regulation are transcription factors (TFs): proteins that can act as activators or repressors of gene expression, typically by binding to regulatory DNA regions and recruiting the transcriptional apparatus (Yamaguchi *et al.*, 2017). TFs are widely used in methods that reconstruct transcriptional networks, and algorithms that reconstruct such networks consider both positive and negative target associations, consistent with mechanistic studies that have

demonstrated the dual-function roles of TFs (Lubelsky and Shaul, 2019). In such a network, each regulator and its target genes form a regulatory unit, or regulon (Margolin *et al.*, 2006). Target genes can belong to multiple regulons, and regulators may co-operate and compete in influencing target gene expression.

In previous studies, we have used regulon activities to identify TFs associated with variant risk loci in breast cancer (Castro *et al.*, 2016), and to characterize differences between molecular subtypes in muscle-invasive bladder cancer (Robertson *et al.*, 2017). Because regulators can co-operate and compete, we anticipated that identifying pairs of regulons that share targets could be informative. Here, we report *RTNduals*,

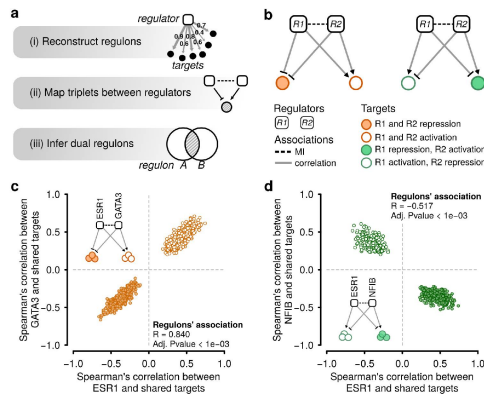


Fig. 1. Inference of dual regulons. (a) *RTNduals* computes dual regulons using: (i) MI between a regulator and targets; (ii) triplets consisting of pairs of regulators and a shared target; (iii) whether the number of shared targets is statistically significant. (b) Examples showing two associated regulators and two regulator-target triples. Left: an example in which the regulators co-operate by influencing shared target genes in the same direction (i.e. either co-activating or co-repressing the shared targets). Right: regulators compete by influencing shared target genes in opposite directions. (c, d) Distribution of correlation coefficients between regulators and shared targets in two example dual regulons, computed from the expression profiles of METABRIC breast cancer data, $n=997$ for cohort 1 (Curtis *et al.*, 2012). For additional details on these examples, please refer to the **Supplementary Information**.

an R/Bioconductor package that automates the search for co-regulation between regulons, assessing all targets shared by pairs of regulators; when it identifies that a pair has more shared targets than expected by chance, which we assess by overlap and permutation analyses, it defines this pair as a *dual regulon*. In the **Supplementary Information** we report two case studies that explore dual regulons in breast cancer TF regulatory networks.

2 A method for identifying dual regulons

Figure 1a gives an overview of how *RTNduals* infers dual regulons. The package can take two types of data as input. The first type consists of a gene expression matrix (e.g. a cancer cohort's transcriptome) and, from prior biological information, a list that indicates which genes should be regarded as regulators. The second consists of a transcriptional regulatory network pre-computed by the *RTN* package (Castro *et al.*, 2016). The package architecture allows the input of different classes of regulators (e.g. TFs, miRNAs).

RTNduals uses three complementary statistics to identify dual regulons (**Fig. 1a**). (i) Targets are assigned to regulators based on mutual information (MI), forming regulons. The statistical significance of the MI values is assessed by permutation and bootstrap analysis. Because regulators can target each other, associations between pairs of regulators are also identified. (ii) Triplets formed by two regulators and a shared target gene are identified, and the direction of regulation is determined by correlation analysis (e.g. Pearson or Spearman). (iii) A Fisher's exact test assesses the number of triplets shared between two regulators, and permutation analysis tests the statistical significance of the correlation between shared targets. The schematics in **Figure 1b** show the two cases that *RTNduals* identifies: regulator pairs (left) that co-operate, influencing shared target genes in the same direction (co-activation or co-repression), and (right)

that compete, influencing targets in opposite directions. **Figure 1c** shows the distribution of Spearman correlations of targets shared between ESR1 and GATA3 regulons, indicating that these TFs either co-activate or co-repress their shared targets, while **Figure 1d** shows a contrasting case with ESR1 and NFIB regulons.

3 Case studies

RTNduals allows high-throughput screening for co-regulators and their shared targets. The **Supplementary Information** provides two detailed case studies that demonstrate the package's workflow, using tumour samples from breast cancer cohorts. The first study analyses a regulatory network generated by *RTN* from METABRIC microarray data (Curtis *et al.*, 2012), while the second case study shows how to prepare harmonized RNA-seq data from the National Cancer Institute's Genomic Data Commons (GDC) for analysis (TCGA, 2012). Dual regulons identified in both case studies are consistent with experimentally supported associations (**Supplementary Tables 1 and 3**), particularly between GATA3, ESR1 and FOXA1 regulons, whose regulators interact physically (Theodorou *et al.*, 2013), all of which are highly influential in ER+ breast cancer.

Acknowledgements

We thank Carolina Mathias and Anderson Scorsato for advice and critical reading the manuscript.

Funding

This work was supported by the National Council for Scientific and Technological Development (CNPq) [407090/2016-9]; and the Cancer Research UK (CRUK), the Breast Cancer Research Foundation (BCRF) [BCRF-17-127]. VSC, CSG, KGO and ST are funded by the Coordination for the Improvement of Higher Education Personnel (CAPES). SJMJ and AGR are funded by the National Cancer Institute of the National Institutes of Health [U24CA210952]. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Castro, M.A.A. *et al.* (2016) Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature Genetics*, **48**(1), 12-21.
- Curtis, C. *et al.* (2012) The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups. *Nature*, **486**(7403), 346-352.
- Fletcher, M.N.C. *et al.* (2013) Master regulators of FGFR2 signalling and breast cancer risk. *Nature Communications*, **4**, 2464.
- Lubelsky, Y.; Shaul, Y. (2019) Recruitment of the protein phosphatase-1 catalytic subunit to promoters by the dual-function transcription factor RFX1. *Biochemical and Biophysical Research Communications*, **509**(4), 1015-1020.
- Margolin, A.A. *et al.* (2006) ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinformatics*, **7**(7), 147-2105.
- Robertson, A.G. *et al.* (2017) Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, **171**(3), 540-560.
- The Cancer Genome Atlas Network (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**(7418), 61-70.
- Theodorou, V. *et al.* (2013) GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Research*, **23**, 12-22.
- Yamaguchi N. *et al.* (2017) Down-regulation of Forkhead box protein A1 (FOXA1) leads to cancer stem cell-like properties in tamoxifen-resistant breast cancer cells through induction of interleukin 6. *Journal of Biological Chemistry*, **292**(20), 8136-8148.

Supplementary Information

***RTN* duals case studies: exploring dual regulons in breast cancer regulatory networks.**

Vinicius S. Chagas^{1,*}, Clarice S. Groeneveld^{1,*}, Kelin G. de Oliveira^{1,2}, Sheyla Trefflich³, Rodrigo C. de Almeida⁴, Bruce A. J. Ponder⁵, Kerstin B. Meyer^{5,6}, Steven J. M. Jones⁷, A. Gordon Robertson^{7,**}, Mauro A. A. Castro^{1,**}.

¹Bioinformatics and Systems Biology Lab, Federal University of Paraná, Curitiba, 81520-260, Brazil. ²Department of Clinical Sciences, Lund University, Lund, 221 85, Sweden. ³Bioinformatics Department, Federal University of Minas Gerais, Belo Horizonte, 31270-901, Brazil. ⁴Department of Medical Statistics and Bioinformatics, Section Molecular Epidemiology, Leiden University Medical Center, Leiden, The Netherlands. ⁵Department of Oncology and Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, United Kingdom. ⁶Wellcome Sanger Institute, Hinxton, CB10 1SA, United Kingdom. ⁷Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, V5Z 4S6, Canada.

*These authors contributed equally to this work.

**Corresponding authors: mauro.castro@ufpr.br, grobertson@bcgsc.ca

Contents

1. METABRIC breast cancer cohort 1	2
1.1 Context	2
1.2 Package installation and data sets	2
1.3 Data preprocessing	2
1.4 A single step infers <i>dual regulons</i>	3
1.5 Representing <i>dual regulons</i> with scatter plots	3
1.6 A heatmap of all tested regulon pairs	4
1.8 Other tools for inferring co-regulation	5
2 TCGA breast invasive carcinoma cohort (TCGA-BRCA)	7
2.1 Context	7
2.2 Using TCGAblinks to download data from GDC	7
2.3 Infer the regulatory network with <i>RTN</i>	8
2.4 Inference of <i>dual regulons</i>	10
2.5 Retrieving results for regulon overlaps	10
Session information	11
Supplementary References	13

1. METABRIC breast cancer cohort 1

1.1 Context

Fletcher *et al.* (2013) reconstructed regulons for 809 transcription factors (TFs) using microarray transcriptomic data from the METABRIC breast cancer cohort (Curtis *et al.*, 2012). Castro *et al.* (2016) found that 36 of these TF regulons were associated with genetic risk of breast cancer. The risk TFs were in two distinct clusters. The “cluster 1” risk TFs were associated with estrogen receptor-positive (ER+) breast cancer risk and comprise TFs such as ESR1, FOXA1, and GATA3, whereas the “cluster 2” risk TFs were associated with estrogen receptor-negative (ER-), basal-like breast cancer. **Our goals here are (1) to explore associations between the regulons reconstructed by Fletcher *et al.* (2013) and (2) to identify statistically inferred *dual regulons*.**

1.2 Package installation and data sets

The *RTNduals* package is available from the R/Bioconductor repository, together with other required packages. Installing and then loading the *Fletcher2013b* data package will make available all data required for this case study.

```
##-- Set the Bioconductor repository
##-- Please make sure to use bioc version >= 3.8 (R >= 3.5)
source("https://bioconductor.org/biocLite.R")
biocVersion()

##-- Install RTNduals and other required packages
##-- RTN(>=2.6.3); RTNduals(>=1.6.2); Fletcher2013b(>=1.16.0)
biocLite(c("RTNduals","Fletcher2013b"))
install.packages("pheatmap")

##-- Call packages
library(RTNduals)
library(Fletcher2013b)

##-- Load 'rtni1st' data object, which includes regulons and expression profiles
data("rtni1st")

##-- A list of transcription factors of interest (here 36 risk-associated TFs)
risk.tfs <- c("AFF3", "AR", "ARNT2", "BRD8", "CBFB", "CEBPB", "E2F2", "E2F3", "ENO1",
             "ESR1", "FOSL1", "FOXA1", "GATA3", "GATAD2A", "LZTFL1", "MTA2", "MYB",
             "MZF1", "NFIB", "PPARD", "RARA", "RB1", "RUNX3", "SNAPC2", "SOX10",
             "SPDEF", "TBX19", "TCEAL1", "TRIM29", "XBP1", "YBX1", "YPEL3", "ZNF24",
             "ZNF434", "ZNF552", "ZNF587")
```

1.3 Data preprocessing

The data preprocessing consists of a single step that creates an **MBR-class** object. This step uses the `tni2mbrPreprocess` function, which requires (1) a transcriptional regulatory network computed by the *RTN* package, and (2) a list of regulators. We will also need to update the `rtni1st` object with the `tni.dpi.filter` function, setting `eps = NA`. As we explain in **section 2.3**, the `eps` argument sets the ARACNe algorithm’s mutual information (MI) threshold. When we want to remove all dependencies between regulons, *e.g.* to enrich regulons with direct targets, we recommend setting `eps = 0.0`, which is the default option. The regulatory network that we’ve loaded would have been calculated with this setting. For *RTNduals*, however, we want to assess the overlap between pairs of target gene sets, so we need to re-run this step on the `rtni1st`

object. We recommend setting `eps = NA`, which will estimate a nonzero MI threshold from the empirical null distribution computed in the permutation and bootstrap steps.

```

#-- Update the 'rtni1st' object
rtni1st <- tni.dpi.filter(rtni1st, eps = NA)

#-- Check consistency of the input data and build an MBR-class object
mbr1st <- tni2mbrPreprocess(tni = rtni1st, regulatoryElements = risk.tfs)

```

1.4 A single step infers *dual regulons*

The `mbrAssociation` function tests the association between pairs of regulons, using Fisher's exact test and permutation analysis to assess the statistical significance of the overlap and the correlation between regulons, respectively.

```

#-- Run 'mbrAssociation' pipeline
mbr1st <- mbrAssociation(mbr1st)

#-- Get a list of dual regulons ranked by correlation statistics
#-- (Supplementary Table 1)
mbr1st_results <- mbrGet(mbr1st, "dualsCorrelation")
mbr1st_results[1:10,]

```

Supplementary Table 1. Top 10 *Dual regulons* ranked by P-value.

Dual Regulon	MI Regulators	R Regulons	Pvalue	Adjusted Pvalue
ESR1~FOXA1	0.34	0.84	2.32e-117	1.46e-114
ESR1~GATA3	0.55	0.84	2.02e-112	1.27e-109
FOXA1~GATA3	0.43	0.81	5.95e-111	3.75e-108
AFF3~ESR1	0.32	0.72	2.06e-85	1.30e-82
MYB~FOXA1	0.30	0.73	7.47e-85	4.71e-82
RUNX3~FOXA1	0.20	-0.73	1.81e-82	1.14e-79
ESR1~MYB	0.38	0.73	8.10e-82	5.10e-79
XBP1~FOXA1	0.54	0.73	7.17e-81	4.52e-78
XBP1~ESR1	0.33	0.68	2.70e-80	1.70e-77
AFF3~GATA3	0.33	0.69	9.11e-78	5.74e-75

*Benjamini-Hochberg (BH) adjusted p-values

Supplementary Table 1 shows the top 10 dual regulons in the `mbr1st_results` object. All but one of them have positive Spearman correlation coefficients (*R Regulons* > 0), meaning that both TFs co-operate by influencing shared target genes in the same direction (*i.e.* either co-activating or co-repressing the shared targets, see below).

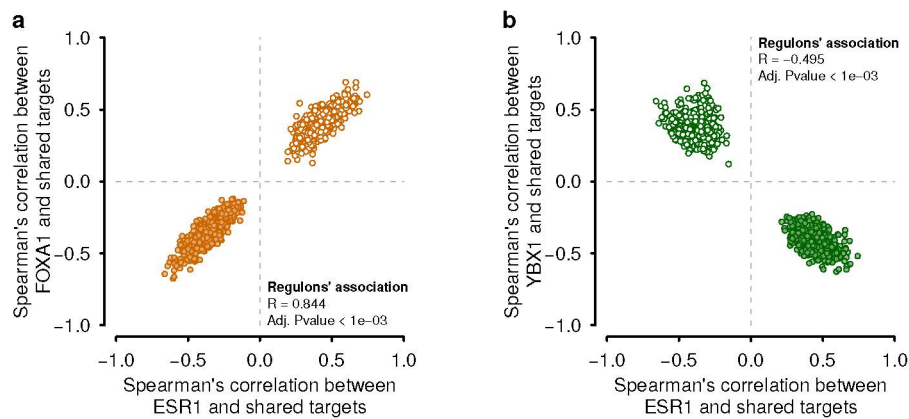
1.5 Representing *dual regulons* with scatter plots

The `mbrPlotDuals` function allows us to represent how the expression of the shared targets of a dual regulon is correlated with the expression of each regulator. In **Supplementary Figure 1** we show *ESR1~FOXA1* and *ESR1~YBX1* as examples generated by this function.

```

#-- Scatter plots of shared targets
#-- (Supplementary Figure 1)
mbrPlotDuals(mbr1st, "ESR1~FOXA1")
mbrPlotDuals(mbr1st, "ESR1~YBX1")

```



Supplementary Figure 1: Relationship between the shared targets in a dual regulon, computed from the expression profiles of METABRIC breast cancer data, $n=997$ (Fletcher *et al.*, 2013). (a) Transcription factors ESR1 and FOXA1 co-operate in influencing shared target genes, while (b) ESR1 and YBX1 have opposite-signed correlations (see **Figure 1c,d** for ESR1~GATA3 and ESR1~NFIB dual regulons, respectively).

1.6 A heatmap of all tested regulon pairs

We can now visualize a heatmap to summarize the relationships between all regulon pairs assessed in the analysis pipeline. For this, we need to call the `mbrGet` function to obtain a correlation matrix with all Spearman correlation coefficients, and then plot a heatmap with the help of the `pheatmap` package.

```

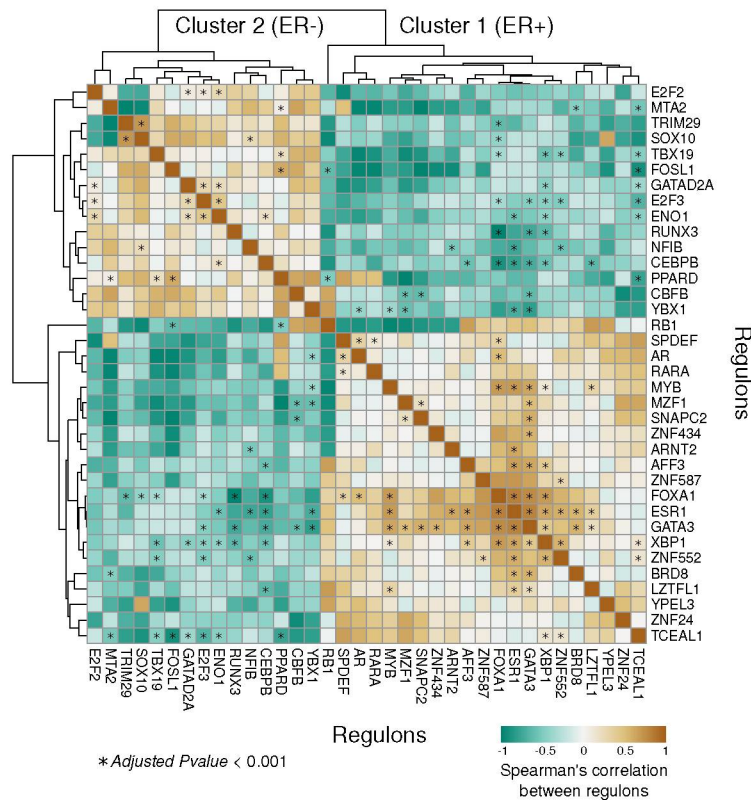
#-- Get the correlation matrix between regulons
dmat <- mbrGet(mbrlst, what="dualsCorMatrix")

#-- Plot the correlation matrix between regulons
#-- (Supplementary Figure 2)
library(pheatmap)
colorpal <- c("#018571", "#80CDC1", "#F5F5F5", "#DFC27D", "#A6611A")
pheatmap(mat = dmat$cormat, display_numbers = dmat$sigmat,
         color = colorRampPalette(colorpal)(100),
         clustering_distance_rows = "correlation",
         clustering_distance_cols = "correlation")

```

Each square in the heatmap of the **Supplementary Figure 2** represents a Spearman correlation coefficient estimated for a regulon pair, and summarizes the more detailed information shown in the scatter plots of the **Supplementary Figure 1** for individual dual regulons. Since the heatmap represents a correlation matrix, values are mirrored across the diagonal. All significant associations ($P < 0.001$, BH adjusted) are marked with asterisks.

The lower right corner of “Cluster 1” contains a region with highly correlated regulons that is enriched with significant predictions, particularly among GATA3, ESR1 and FOXA1, all of which are highly influential in ER+ breast cancer (Theodorou *et al.*, 2013). Duals *ESR1~YBX1* (**Supplementary Figure 1b**) and *ESR1~NFIB* (**Figure 1d**) are also of note: both NFIB and YBX1 interact with the ESR1-FOXA1 complex and inhibit the transactivational potential of ESR1, and these interactions further repress ESR1 target gene expression when in association with induced FGFR2 signalling (Campbell *et al.*, 2018). There is also evidence linking SOX10 and TRIM29 (Panaccione *et al.*, 2017) (as indicated in **Supplementary Figure 2**), both of which are associated with a neural stem cell-like signature in ER- tumours.



Supplementary Figure 2: Heatmap showing the correlation matrix between regulons for 36 transcription factors. Each point in the heatmap summarizes the relationship between a regulon's shared targets as shown in the scatter plots of **Supplementary Figure 1**. Significant associations ($P < 0.001$, BH adjusted) are indicated with asterisks. "Cluster 1" and "Cluster 2", as named in Castro et al. (2016), represent regulons associated with ER+ and ER- tumours, respectively.

1.8 Other tools for inferring co-regulation

Several algorithms have been developed to explore regulation in biological networks. *Bionet* (Beisser et al., 2010), *Minet* (Meyer et al., 2008), and *CoRegNet* (Nicolle et al., 2015) are examples of R/Bioconductor packages that implement routines and hypothesis testing methods for functional analysis of biological networks. Nicolle et al. (2015) compiled features of a large number of other tools to compare state of the art methods and concluded that, by that time, only *CoRegNet* provided methods to infer co-regulation. The same authors assessed *RTN* as able to do "network inference", "genomic data integration", and "differential analysis". Now with *RTNduals* we can further contextualize the commonalities between these tools. Next we run the *CoRegNet* pipeline with the same gene expression data used in *RTNduals*.

```

#-- Run CoRegNet with the same input data used in RTNduals
library(CoRegNet)
gexp1st <- tni.get(rtnei1st, what="gexp", idkey = "SYMBOL")
coreg1st <- hLICORN(gexp1st, TFlist=risk.tfs)

#-- Get co-regulators ranked by CoRegNet 'support' statistics
#-- (Supplementary Table 3)
coreg1st_results <- coregulators(coreg1st)
coreg1st_results[1:10,]

```

Supplementary Table 3. Top 10 co-regulators supported by *CoRegNet*.

Co-regulator	Reg1	Reg2	Support	Fisher Test	Adjusted Pvalue
ESR1~GATA3**	ESR1	GATA3	0.053	0.00e+00	0.00e+00
ESR1~AFF3**	ESR1	AFF3	0.041	0.00e+00	0.00e+00
AFF3~GATA3**	AFF3	GATA3	0.037	0.00e+00	0.00e+00
ESR1~FOXA1**	ESR1	FOXA1	0.032	6.87e-124	6.38e-123
FOXA1~GATA3**	FOXA1	GATA3	0.029	1.42e-41	5.61e-41
ESR1~MYB**	ESR1	MYB	0.024	2.38e-86	1.50e-85
FOXA1~AFF3	FOXA1	AFF3	0.019	2.57e-46	1.10e-45
MYB~GATA3**	MYB	GATA3	0.018	8.85e-11	2.59e-10
ESR1~XBP1**	ESR1	XBP1	0.016	8.36e-257	1.47e-255
AFF3~XBP1**	AFF3	XBP1	0.014	7.76e-266	1.53e-264

**Co-regulators also listed as dual regulons in Supplementary Figure 2.

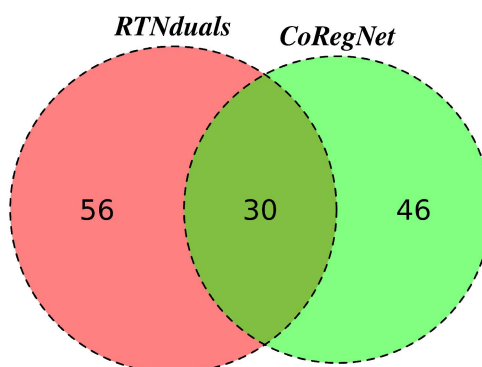
Supplementary Table 3 shows that the top 10 co-regulators supported by *CoRegNet* are consistent with the regulons identified by *RTNduals*, with nine of them listed in **Supplementary Figure 2**. Next we extend the comparison with *RTNduals* to all co-regulators.

```

#-- Match labels between CoRegNet and RTNduals
lab <- paste(coreg1st_results$Reg1, coreg1st_results$Reg2, sep="~")
idx <- !(lab %in% rownames(mbr1st_results))
lab[idx] <- paste(coreg1st_results$Reg2, coreg1st_results$Reg1, sep="~")[idx]
rownames(coreg1st_results) <- lab

#-- Plot a Venn diagram (Supplementary Figure 3)
library(VennDiagram)
res <- list(RTNduals = rownames(mbr1st_results),
           CoRegNet = rownames(coreg1st_results))
venn.diagram(res, fill=c("red","green"), alpha=c(0.5,0.5), cex=2, cat.cex=2,
             cat.pos=0, cat.fontface=4, lty=2, fontfamily=3, filename="venn.tiff")

```



Supplementary Figure 3: Venn diagram showing the overlap between dual regulons and co-regulators (inferred by *RTNduals* and *CoRegNet*, respectively) for the same input gene expression data.

Supplementary Figure 3 shows that 1/3 of the co-regulators are also identified as dual regulons, and the differences might be explained by the input regulatory networks, which are not the same. These packages use different algorithms to compute the regulatory networks; *CoRegNet* detects TF-gene interactions with the LICORN algorithm (Rouveirol *et al.*, 2007), while *RTN* reconstructs regulons with the ARACNe algorithm (Margolin *et al.*, 2006) (additional comments in **section 2.3**). After computing the regulatory network, both packages use similar approaches to access the number of targets shared between a pair of regulators. The main conceptual difference relies on what is considered a regulatory unit, which shapes the analysis workflows. For *RTNduals*, the regulatory unit is the regulon (*e.g.* group of genes and a given regulator) and regulation is investigated between regulons. For *CoRegNet*, the regulatory unit is formed by one gene and two regulators (*e.g.* cooperative regulations are enumerated in the first place) and regulation is investigated between regulators. Therefore, these packages take complementary directions: *RTNduals* implements a top-down approach, breaking down regulons to gain insight into the subsystems, while *CoRegNet* implements a bottom-up strategy, linking together individual co-regulatory elements to form larger subsystems. Users might benefit exploring regulatory associations with both packages to check the overall consistency of the results.

2 TCGA breast invasive carcinoma cohort (TCGA-BRCA)

2.1 Context

In **section 1**, we used a precalculated transcriptional network for the METABRIC breast cancer cohort, which we made available as the **Fletcher2013b** data package. In **section 2**, we will show how to prepare input data for the *RTN* and *RTNduals* packages using the publicly available mRNA-seq data for the TCGA-BRCA cohort. We will show how to download harmonized GRCh38/hg38 data from the Genomic Data Commons (GDC) using the TCGAbiolinks package (Colaprico *et al.*, 2016). The preprocessing generates a **SummarizedExperiment** object that contains gene expression data, which is then used to generate the transcriptional network. The subsequent steps will infer *dual regulons* following exactly the same steps as described in **section 1.4**.

2.2 Using TCGAbiolinks to download data from GDC

Please make sure you have all libraries installed before proceeding.

```
library(SummarizedExperiment)
library(TCGAbiolinks)
library(TxDb.Hsapiens.UCSC.hg38.knownGene)
```

We'll use the Bioconductor package **TCGAbiolinks** to query and download from GDC. We are looking for the harmonized, pre-processed RNA-seq for the TCGA-BRCA cohort.

TCGAbiolinks will create a directory called **GDCdata** in your working directory and will save into it the files downloaded from GDC. The download can take a while. The files for each patient will be downloaded in a separate file. Then, the **GDCprepare** function will compile them into an R object of class **RangedSummarizedExperiment**.

The **RangedSummarizedExperiment** has 6 slots. The most important are **rowRanges** (gene metadata), **colData** (patient metadata), and **assays**, which contains the gene expression matrix.

```
##-- Subset BRCA cohort for a quicker demonstration
subsample <- TCGAquery_subtype(tumor = "BRCA")
subsample <- subsample$patient[sample(nrow(subsample), 500)]
```

```

#-- Download mRNA TCGA-BRCA data
query <- GDCquery(project= "TCGA-BRCA",
  data.category = "Transcriptome Profiling",
  data.type = "Gene Expression Quantification",
  experimental.strategy = "RNA-Seq",
  workflow.type = "HTSeq - FPKM",
  sample.type = c("Primary solid Tumor"),
  barcode = subsample)

GDCdownload(query)
tcgaBRCA_mRNA_data <- GDCprepare(query)

```

The object downloaded from GDC contains gene-level expression data that includes both coding and noncoding genes (*e.g.* lincRNAs). We will filter these, retaining only genes annotated in the UCSC hg38 known gene list.

```

#-- Subset by known gene locations
geneRanges <- genes(TxDb.Hsapiens.UCSC.hg38.knownGene)
tcgaBRCA_mRNA_data <- subsetByOverlaps(tcgaBRCA_mRNA_data, geneRanges)

```

Finally, we'll change column names for better internal pre-processing in RTN's `tni.constructor` function. Having the `SYMBOL` column will enable genes with the same symbol to be preprocessed by this function. The `tcgaBRCA_mRNA_data` object is ready for the RTN pipeline.

```

#-- Change column names for best 'tni.constructor' summarizations
#-- and save the preprocessed data for subsequent analyses
colnames(rowData(tcgaBRCA_mRNA_data)) <- c("ENSEMBL", "SYMBOL", "OG_ENSEMBL")
save(tcgaBRCA_mRNA_data, file = "tcgaBRCA_mRNA_data_preprocessed.RData")

```

2.3 Infer the regulatory network with *RTN*

The RTN pipeline starts with the construction of a `TNI-class` object, using the `tni.constructor` method. This method takes in a matrix of gene expression and metadata on the samples and genes, as well as a list of the regulators to be evaluated. Here, the expression matrix and metadata are available as a `SummarizedExperiment` object, and the list of regulators will be extracted from the `rtni1st` object with the `tni.get` accessory function. The `tni.constructor` method will check the consistency of all the given arguments. The inference pipeline is then executed in three subsequent steps: (i) compute mutual information (MI) between a regulator and all potential targets, removing non-significant associations by permutation analysis, (ii) remove unstable interactions by bootstrap, and (iii) apply the ARACNe algorithm, which uses the data processing inequality (DPI) theorem to remove indirect interactions (for additional details, please refer to Margolin *et al.* (2006) and Fletcher *et al.* (2013)). Briefly, consider three random variables, **X**, **Y** and **Z** forming a network triplet, with **X** interacting with **Z** only through **Y** (*i.e.*, the interaction network is **X**->**Y**->**Z**), and no alternative path exists between **X** and **Z**). The DPI theorem states that the information transferred between **Y** and **Z** is always larger than the information transferred between **X** and **Z**. Based on this assumption, the ARACNe algorithm scans all triplets formed by two regulators and one target and removes the edge with the smallest MI value of each triplet, which is regarded as a redundant association. As the dependencies between regulators are eliminated in the DPI filter, the overlap between regulons is observed in the interactions not removed by the ARACNe algorithm.

```

#-- Get the list of regulatoryElements available from the 'rtni1st' object
regulatoryElements <- names(tni.get(rtni1st, what="regulatoryElements"))

#-- TNI constructor
rtni_tcgaBRCA <- tni.constructor(tcgaBRCA_mRNA_data,
  regulatoryElements = regulatoryElements)

```

To compute a large regulatory network we recommend using a multithreaded mode with the **snow** package. As minimum computational resources, we suggest a processor with ≥ 4 cores and RAM ≥ 8 GB per core (specific routines should be adjusted for the available resources). The **makeCluster** function will set the number of nodes to create on the local machine, making a **cluster** object available for the **TNI-class** methods. This example (29885 rows vs. 500 columns gene expression matrix and 809 regulators) should take 2h to conclude when running in a Core i9-8950H workstation with 32GB DDR4 RAM.

```
##-- Compute the reference regulatory network by permutation and
##-- bootstrap analyses. For RNA-seq data we recommend using the
##-- non-parametric estimator of the mutual information
##-- (estimator = "spearman").
library(snow)
options(cluster=makeCluster(4, "SOCK"))
rtni_tcgaBRCA <- tni.permutation(rtni_tcgaBRCA, pValueCutoff = 10^-7,
                               estimator = "spearman")
rtni_tcgaBRCA <- tni.bootstrap(rtni_tcgaBRCA, nBootstraps = 200)
stopCluster(getOption("cluster"))
```

Next we run the ARACNe algorithm. Note that the overlap between regulons is affected by the **eps** argument, which sets the threshold for removing the edge with the smallest MI value of each triplet (see comments above and the **aracne** function). In order to access the overlap between regulons in *RTNduals*, we recommend setting **eps = NA**, which will estimate the MI threshold from the empirical null distribution computed in the permutation and bootstrap steps.

```
##-- Compute the DPI-filtered regulatory network
rtni_tcgaBRCA <- tni.dpi.filter(rtni_tcgaBRCA, eps = NA)

##-- Save the TNI object for subsequent analyses
save(rtni_tcgaBRCA, file="rtni_tcgaBRCA.RData")
```

Please note that some level of missing annotation is expected, as not all gene symbols listed in the **regulatoryElements** might be available in the TCGA-BRCA preprocessed data. Also, inconsistent data might be removed in the **tni.constructor** preprocess; for example, it is not possible to test associations for a gene whose expression does not vary across samples, so this gene is not included in the analysis (in this example, genes “SOX10”, “XBP1” and “ZNF434” were missed or removed, so we tested fewer regulons than those tested in **section 1**). For a summary of the resulting regulatory network we recommend using the **tni.regulon.summary** function.

```
##-- Summary
tni.regulon.summary(rtni_tcgaBRCA)
```

```
## This regulatory network comprised of 771 regulons.
```

```
## -- DPI-filtered network:
```

```
## regulatoryElements      Targets      Edges
##           771           16754      74981
##   Min. 1st Qu.  Median   Mean 3rd Qu.  Max.
##     0.0   14.0   63.0   97.3  135.0  1218.0
```

```
## -- Reference network:
```

```
## regulatoryElements      Targets      Edges
##           771           16754      565928
##   Min. 1st Qu.  Median   Mean 3rd Qu.  Max.
##     0.0   31.5  341.0   734.0 1111.0  4380.0
## ---
```

Additionally, all parameters, input data, and results available in the final `TNI-class` object can be retrieved by the `tni.get` accessory function.

```
##-- For example, to retrieve DPI-filtered regulons with a given annotation
regulons <- tni.get(rtni_tcgaBRCA, what="regulons", idkey="SYMBOL")
```

Note that the MI statistics are based on a gene's expression varying across a cohort. If a gene's expression does not vary across a cohort, it is not possible to associate this gene's expression with the expression of other genes in the cohort. As an extreme case, genes that exhibit no variability (*e.g.* that are not expressed in all samples) are excluded from the analysis. Large cohorts of tumour samples typically contain multiple molecular subtypes, and typically provide good expression variability for building regulons. In contrast, sample sets that are more homogeneous may be more challenging to explore with regulons, and this may be the case with sets of normal, non-cancerous samples. We do not recommend computing regulons for cohorts of low variability, or for subsets of a cohort.

2.4 Inference of *dual regulons*

The `mbrAssociation` function will call *dual regulons* following exactly the same steps as described in [section 1.4](#), but now for regulons computed for the TCGA-BRCA cohort.

```
##-- Run 'tni2mbrPreprocess' and check datasets
mbr_tcgaBRCA <- tni2mbrPreprocess(tni = rtni_tcgaBRCA, regulatoryElements = risk.tfs)
```

```
##-- Run 'mbrAssociation' pipeline
mbr_tcgaBRCA <- mbrAssociation(mbr_tcgaBRCA)
```

```
##-- Get a list of dual regulons
##-- (Supplementary Table 4)
mbr_tcgaBRCA_results <- mbrGet(mbr_tcgaBRCA, "dualsCorrelation")
mbr_tcgaBRCA_results[1:10,]
```

Supplementary Table 4. Top 10 *dual regulons* in the TCGA-BRCA cohort.

Dual Regulon	MI Regulators	R Regulons	Pvalue	Adjusted Pvalue
ESR1~FOXA1**	0.38	0.77	1.73e-83	8.59e-81
FOXA1~GATA3**	0.45	0.70	8.44e-79	4.19e-76
CEBPB~FOXA1**	0.24	-0.65	1.45e-64	7.17e-62
FOXA1~YBX1	0.27	-0.57	1.53e-58	7.57e-56
ESR1~GATA3**	0.35	0.71	2.07e-58	1.03e-55
AR~FOXA1**	0.23	0.60	2.00e-52	9.93e-50
AFF3~ESR1**	0.29	0.64	5.23e-50	2.59e-47
ESR1~ZNF552**	0.32	0.74	4.30e-46	2.13e-43
GATA3~TCEAL1	0.32	0.71	2.79e-43	1.39e-40
ESR1~YBX1**	0.31	-0.57	1.03e-41	5.10e-39

*Benjamini-Hochberg (BH) adjusted p-values

**Dual regulons also listed in Supplementary Figure 2

Supplementary Table 4 shows the top 10 dual regulons in the `mbr_tcgaBRCA_results` object; nine of these are listed in **Supplementary Figure 2**, which shows dual regulons inferred in the microarray-based METABRIC cohort 1 data.

2.5 Retrieving results for regulon overlaps

Next we show how to retrieve information for a dual regulon from an `MBR-class` object, including the genes in each regulon. We also demonstrate how the overlap counts are obtained between two regulons.


```

##-- Get a list of dual regulons ranked by the overlap statistics
mbr_tcgaBRCA_overlap <- mbrGet(mbr_tcgaBRCA, "dualsOverlap")
mbr_tcgaBRCA_overlap[1:3,-c(1,2)]

##          Universe.Size Regulon1.Size Regulon2.Size Expected.Overlap
## FOXA1~GATA3          16754           797           406          19.313716
## NFIB~TBX19           16754           222           279           3.696908
## ESR1~GATA3           16754           341           406           8.263459
##          Observed.Overlap          Pvalue Adjusted.Pvalue
## FOXA1~GATA3           124 7.594822e-67  3.767032e-64
## NFIB~TBX19            66 4.138657e-65  2.052774e-62
## ESR1~GATA3            63 1.542607e-37  7.651332e-35

##-- Extract regulons and network summary
tni <- mbrGet(mbr_tcgaBRCA, what="TNI")
regulons <- tni.get(tni, what="regulons", idkey = "SYMBOL")
tniSummary <- tni.get(tni, what="summary")

##-- Get the relevant counts to compute the overlap between two
##-- regulons (e.g. FOXA1 and GATA3 regulons).
Universe.Size <- tniSummary$results$tnet[1,"Targets"]
Regulon1.Size <- length(regulons$FOXA1)
Regulon2.Size <- length(regulons$GATA3)
Observed.Overlap <- length(intersect(regulons$FOXA1,regulons$GATA3))

##-- Combine results
c(Universe.Size,Regulon1.Size,Regulon2.Size,Observed.Overlap)

## [1] 16754 797 406 124

```

Session information

```

## R version 3.5.3 (2019-03-11)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.2 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## attached base packages:
## [1] parallel stats4 stats graphics grDevices utils datasets
## [8] methods base
##
## other attached packages:
## [1] TxDb.Hsapiens.UCSC.hg38.knownGene_3.4.0
## [2] GenomicFeatures_1.34.1
## [3] AnnotationDbi_1.44.0
## [4] TCGAbiolinks_2.10.0
## [5] SummarizedExperiment_1.12.0
## [6] DelayedArray_0.8.0
## [7] BiocParallel_1.16.4
## [8] matrixStats_0.54.0

```

```

## [9] Biobase_2.42.0
## [10] GenomicRanges_1.34.0
## [11] GenomeInfoDb_1.18.1
## [12] IRanges_2.16.0
## [13] S4Vectors_0.20.1
## [14] BiocGenerics_0.28.0
## [15] pheatmap_1.0.10
## [16] RTNduals_1.7.2
## [17] Fletcher2013b_1.18.0
## [18] igraph_1.2.2
## [19] RedeR_1.30.0
## [20] RTN_2.7.3
## [21] Fletcher2013a_1.18.0
## [22] limma_3.38.3
##
## loaded via a namespace (and not attached):
## [1] backports_1.1.3          snow_0.4-3
## [3] circlize_0.4.5          aroma.light_3.12.0
## [5] plyr_1.8.4              selectr_0.4-1
## [7] ConsensusClusterPlus_1.46.0 lazyeval_0.2.1
## [9] splines_3.5.3           ggplot2_3.1.0
## [11] sva_3.30.0              digest_0.6.18
## [13] foreach_1.4.4           htmltools_0.3.6
## [15] gdata_2.18.0            magrittr_1.5
## [17] memoise_1.1.0           cluster_2.0.7-1
## [19] doParallel_1.0.14       mixtools_1.1.0
## [21] ComplexHeatmap_1.20.0   Biostrings_2.50.1
## [23] readr_1.3.1             annotate_1.60.0
## [25] R.utils_2.7.0           prettyunits_1.0.2
## [27] colorspace_1.3-2        blob_1.1.1
## [29] rvest_0.3.2             ggrepel_0.8.0
## [31] xfun_0.4                dplyr_0.7.8
## [33] crayon_1.3.4            RCurl_1.95-4.11
## [35] jsonlite_1.6            genefilter_1.64.0
## [37] bindr_0.1.1             zoo_1.8-4
## [39] survival_2.43-3         iterators_1.0.10
## [41] glue_1.3.0              survminer_0.4.3
## [43] gtable_0.2.0            zlibbioc_1.28.0
## [45] XVector_0.22.0          GetoptLong_0.1.7
## [47] shape_1.4.4             scales_1.0.0
## [49] DESeq_1.34.0            futile.options_1.0.1
## [51] DBI_1.0.0               edgeR_3.24.3
## [53] ggthemes_4.0.1         Rcpp_1.0.0
## [55] cmprsk_2.2-7           xtable_1.8-3
## [57] progress_1.2.0         bit_1.1-14
## [59] matlab_1.0.2           km.ci_0.5-2
## [61] httr_1.4.0             gplots_3.0.1
## [63] RColorBrewer_1.1-2     pkgconfig_2.0.2
## [65] XML_3.98-1.16          R.methodsS3_1.7.1
## [67] locfit_1.5-9.1         tidysselect_0.2.5
## [69] rlang_0.3.0.1          munsell_0.5.0
## [71] tools_3.5.3            downloader_0.4
## [73] generics_0.0.2         RSQLite_2.1.1
## [75] broom_0.5.1            evaluate_0.12

```

```

## [77] stringr_1.3.1          yaml_2.2.0
## [79] knitr_1.21             bit64_0.9-7
## [81] survMisc_0.5.5        caTools_1.17.1.1
## [83] purrr_0.2.5           bindrcpp_0.2.2
## [85] nlme_3.1-137          EDASeq_2.16.0
## [87] formatR_1.5           R.oo_1.22.0
## [89] xml2_1.2.0            biomaRt_2.38.0
## [91] compiler_3.5.3        e1071_1.7-0
## [93] minet_3.40.0          viper_1.16.0
## [95] tibble_1.4.2          geneplotter_1.60.0
## [97] stringi_1.2.4         futile.logger_1.4.3
## [99] lattice_0.20-38      Matrix_1.2-17
## [101] KMSurv_0.1-5         pillar_1.3.1
## [103] GlobalOptions_0.1.0  data.table_1.11.8
## [105] bitops_1.0-6         rtracklayer_1.42.1
## [107] R6_2.3.0             latticeExtra_0.6-28
## [109] hwriter_1.3.2        ShortRead_1.40.0
## [111] KernSmooth_2.23-15   gridExtra_2.3
## [113] codetools_0.2-16     lambda.r_1.2.3
## [115] MASS_7.3-51.1        gtools_3.8.1
## [117] assertthat_0.2.0     rjson_0.2.20
## [119] GenomicAlignments_1.18.0 Rsamtools_1.34.0
## [121] GenomeInfoDbData_1.2.0 mgcv_1.8-28
## [123] hms_0.4.2            VennDiagram_1.6.20
## [125] grid_3.5.3           tidyr_0.8.2
## [127] class_7.3-15         rmarkdown_1.11
## [129] segmented_0.5-3.0    ggpubr_0.2

```

Supplementary References

- Beisser,D. *et al.* (2010) BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics*, **26**, 1129–1130.
- Campbell,T.N. *et al.* (2018) ER α Binding by Transcription Factors NFIB and YBX1 Enables FGFR2 Signaling to Modulate Estrogen Responsiveness in Breast Cancer. *Cancer Research*, **78**, 410–421.
- Castro,M.A.A. *et al.* (2016) Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature Genetics*, **48**, 12–21.
- Colaprico,A. *et al.* (2016) TCGAbiolinks: An r/bioconductor package for integrative analysis of tcga data. *Nucleic Acids Research*, **44**, e71.
- Curtis,C. *et al.* (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.
- Fletcher,M.N. *et al.* (2013) Master regulators of FGFR2 signalling and breast cancer risk. *Nature Communications*, **4**, 2464.
- Margolin,A.A. *et al.* (2006) ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
- Meyer,P.E. *et al.* (2008) Minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.
- Nicolle,R. *et al.* (2015) CoRegNet: reconstruction and integrated analysis of co-regulatory networks.

Bioinformatics, **31**, 3066–3068.

Panaccione,A. *et al.* (2017) Expression Profiling of Clinical Specimens Supports the Existence of Neural Progenitor-Like Stem Cells in Basal Breast Cancers. *Clinical Breast Cancer*, **17**, 298–306.e7.

Rouveirol,C. *et al.* (2007) LICORN: learning cooperative regulation networks from gene expression data. *Bioinformatics*, **23**, 2407–2414.

Theodorou,V. *et al.* (2013) GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Research*, **23**, 12–22.

**ANEXO III - THE CONSENSUS MOLECULAR CLASSIFICATION OF MUSCLE-INVASIVE
BLADDER CANCER**

The consensus molecular classification of muscle-invasive bladder cancer

Authors

Aurélien Kamoun^{1§}, Aurélien de Reyniès^{1*}, Yves Allory^{2*}, Gottfrid Sjö Dahl^{3*}, A. Gordon Robertson^{4*}, Roland Seiler⁵, Katherine A. Hoadley⁶, Hikmat Al-Ahmadie⁷, Woonyoung Choi⁸, Clarice S. Groeneveld⁹, Mauro A. A. Castro⁹, Jacqueline Fontugne², Pontus Eriksson¹⁰, Qianxing Mo¹¹, Jordan Kardos⁶, Alexandre Zlotta¹², Arndt Hartmann¹³, Colin P. Dinney¹⁴, Joaquim Bellmunt¹⁵, Thomas Powles¹⁶, Núria Malats¹⁷, Keith S. Chan¹⁸, William Y. Kim¹⁹, David J. McConkey²⁰, Peter C. Black²¹, Lars Dyrskjöt²², Mattias Höglund¹⁰, Seth P. Lerner²³, Francisco X. Real²⁴, François Radvanyi²⁵, The Bladder Cancer Molecular Taxonomy Group⁺

§ corresponding author, *these authors contributed equally to this work

+ A full list of all consortium members appears at the end of the article

1. Cartes d'Identité des Tumeurs Program, French League Against Cancer, Paris, France
2. Department of Pathology, Institut Curie Hospital Group, Paris, France
3. Division of Urological Research, Department of Translational Medicine, Lund University, Skåne University Hospital Malmö, Sweden
4. Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, Canada
5. Department of Urology, Bern University Hospital, Switzerland
6. Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
7. Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA
8. Johns Hopkins Greenberg Bladder Cancer Institute and Brady Urological Institute, Johns Hopkins University, Baltimore, MD, USA
9. Bioinformatics and Systems Biology Laboratory, Federal University of Paraná, Polytechnic Center, Curitiba, Brazil
10. Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden
11. Department of Medicine, Baylor College of Medicine, Houston, TX, USA
12. Department of Surgery, Division of Urology, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, ON, Canada
13. Institute of Pathology, University Erlangen-Nürnberg, Krankenhausstr 8-10, Erlangen, Germany
14. Department of Urology and Department of Cancer Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA
15. Bladder Cancer Center, Dana-Farber/Brigham and Women's Cancer Center, Harvard Medical School, Boston, MA, USA
16. Barts Cancer Institute ECMC, Barts Health and the Royal Free NHS Trust, Queen Mary University of London, London, UK
17. Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
18. Molecular & Cellular Biology/Scott Department of Urology, Baylor College of Medicine, One Baylor Plaza, Houston, TX, USA
19. Department of Genetics, Department of Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

20. Johns Hopkins Greenberg Bladder Cancer Institute and Brady Urological Institute, Johns Hopkins University, Baltimore, MD, USA
21. Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada
22. Department of Molecular Medicine, Aarhus University Hospital, Aarhus 8200, Denmark
23. Scott Department of Urology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA
24. Epithelial Carcinogenesis Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
25. Molecular Oncology, CNRS UMR 144, Institut Curie, Paris, France

Abstract

Muscle-Invasive Bladder Cancer (MIBC) is a molecularly diverse disease with heterogeneous clinical outcomes. Several molecular classifications have been proposed, yielding diverse sets of subtypes, which hampers the clinical implications of such knowledge. Here, we report the results of a large international effort to reach a consensus on MIBC molecular subtypes. Using 1750 MIBC transcriptomes and a network-based analysis of six independent MIBC classification systems, we identified a consensus set of six molecular classes: Luminal Papillary (24%), Luminal Non-Specified (8%), Luminal Unstable (15%), Stroma-rich (15%), Basal/Squamous (35%), and Neuroendocrine-like (3%). These consensus classes differ regarding underlying oncogenic mechanisms, infiltration by immune and stromal cells, and histological and clinical characteristics. This consensus system offers a robust framework that will enable testing and validating predictive biomarkers in future clinical trials.

Bladder cancer is one of the most frequently diagnosed cancers in North America and Europe (4th in men and 9th in women). Most bladder cancers are urothelial carcinoma, which are classified for operational reasons as either non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). MIBC is usually diagnosed *de novo*, but may arise from the 10 to 20% of NMIBC cases that eventually progress. MIBC is the most aggressive disease state and is associated with a five-year survival rate of 60% for patients with localized disease, and less than 10% for patients with distant metastases.

At the molecular level, MIBC is a heterogeneous disease that is characterized by genomic instability and a high mutation rate. Many chromosomal rearrangements and more than 50 oncogenes and tumour suppressor genes have been identified as recurrently altered^{1,2}. Transcriptomic profiling facilitates stratifying bladder cancer into molecular subtypes in order to more precisely classify a patient's cancer according to prognosis and therapeutic options. Various teams have been working on the molecular stratification of bladder cancers, and several expression-based classification schemes have been proposed, either considering the full spectrum of bladder cancers³⁻⁶ or focusing separately on MIBC^{2,7-13} or on NMIBC¹⁴. These classifications have considerably advanced our understanding of bladder cancer biology; for example, the association between molecular subtypes and urothelial differentiation, and similarities between subtypes in bladder cancer and other cancers. In addition, specific genomic alterations were found to be enriched in particular molecular subtypes, including mutations targeting genes involved in cell cycle regulation, chromatin remodelling and receptor tyrosine kinase signaling. Importantly, several reports have highlighted the clinical importance of MIBC molecular stratification, suggesting that responses to chemotherapy and immunotherapy may be enriched in specific MIBC subtypes^{9,15-17}.

The published MIBC classifications share many characteristics, including subtype-specific molecular features; however, the classifications are diverse, containing between two to seven molecular subtypes, and having both shared and unique subtype names. This diversity has hampered transferring subtypes into clinical practice, and highlights that identifying a single set of consensus molecular subtypes would facilitate work to achieve such a transfer.

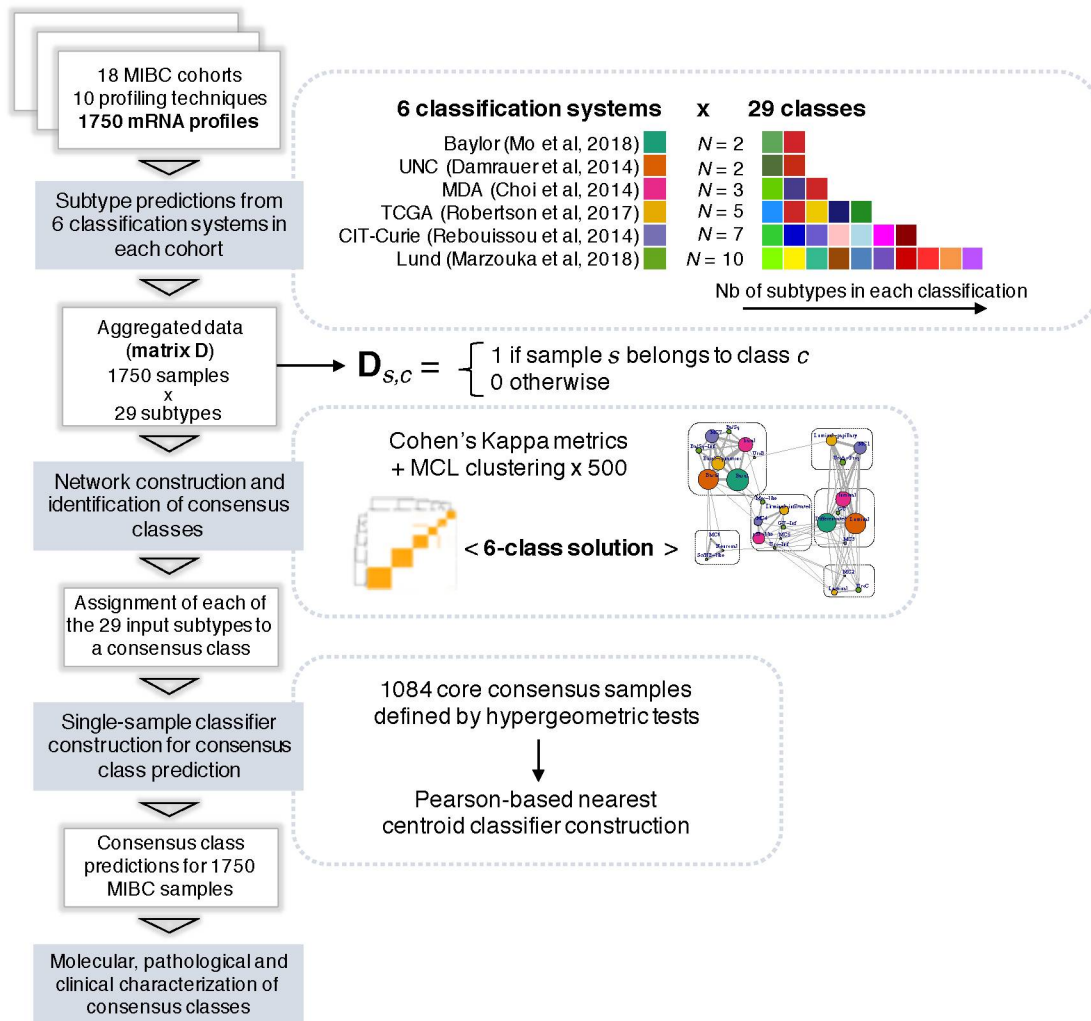
Here, we report the results of an international collaborative effort to reconcile molecular MIBC classifications, involving pathologists, urologists, oncologists, biologists, and bioinformaticians. By analysing six previously published classification schemes and combining public transcriptome data for 1750 tumours, we established a six-class, consensus molecular classification for MIBC. We characterized the consensus classes using additional molecular, pathological and clinical data. To support the use of this consensus molecular classification, we offer a freely available transcriptomic classifier that assigns consensus class labels to single tumour samples (<https://github.com/cit-bioinfo/consensusMIBC>).

Results

Published molecular classifications of MIBC converge on six classes.

We used six published MIBC molecular classifications to define a unified consensus subtyping system. We refer to these input classifications as Baylor (Tumour differentiation)¹³, UNC⁷, CIT-Curie⁸, MDA⁹, Lund¹⁰, and TCGA². Following the approach outlined in Extended data figure 1, we selected 18 MIBC mRNA datasets (n=1750, Supplementary Table 1), and assigned each sample to a subtype in each of the six classification systems.

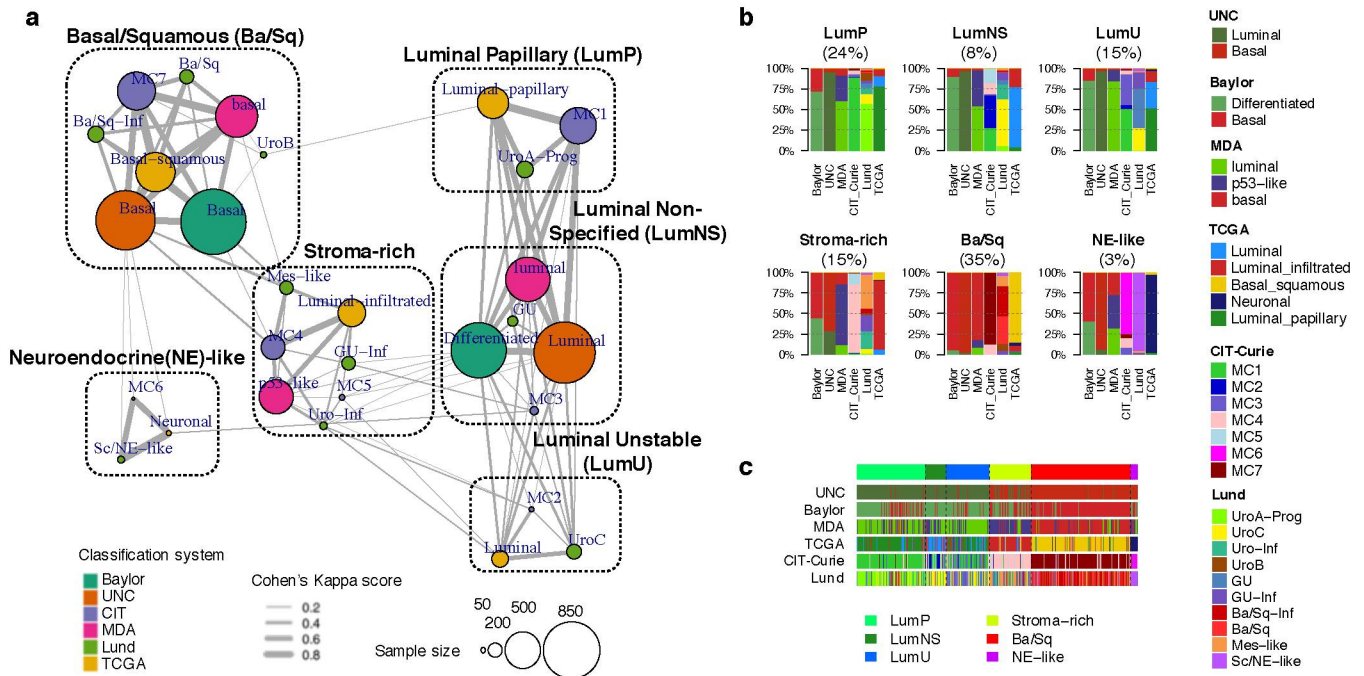
Extended data figure 1 : Analytical workflow



We built a weighted network of these input subtypes, using Cohen's Kappa metric to quantify similarities between subtypes from different classification systems, and applied a Markov cluster clustering algorithm (MCL) to identify robust network substructures corresponding to potential consensus classes (Methods, Supplementary Figure 1). We identified a 6-cluster solution, defining six biologically relevant consensus molecular classes, which we labeled as: Luminal Papillary (LumP), Luminal Non-Specified (LumNS), Luminal Unstable (LumU), Stroma-rich, Basal/Squamous

(Ba/Sq), and Neuroendocrine-like (NE-like) (Figure 1a). Considerations motivating our choices for these consensus names are detailed in the Supplementary Note.

Figure 1 : The six consensus classes and their relation to input molecular subtypes.



The six molecular classes had variable sample sizes, with Ba/Sq and LumP being the most prevalent (35% and 24% of all samples, respectively). The remaining 41% of samples were split into LumU (15%), Stroma-rich (15%), LumN (8%), and NE-like (3%) tumours (Figure 1b). The consensus classification was strongly associated with each of the initial classification systems (Chi-square $P < 10^{-165}$), as illustrated in Figure 1 and Supplementary Figure 2.

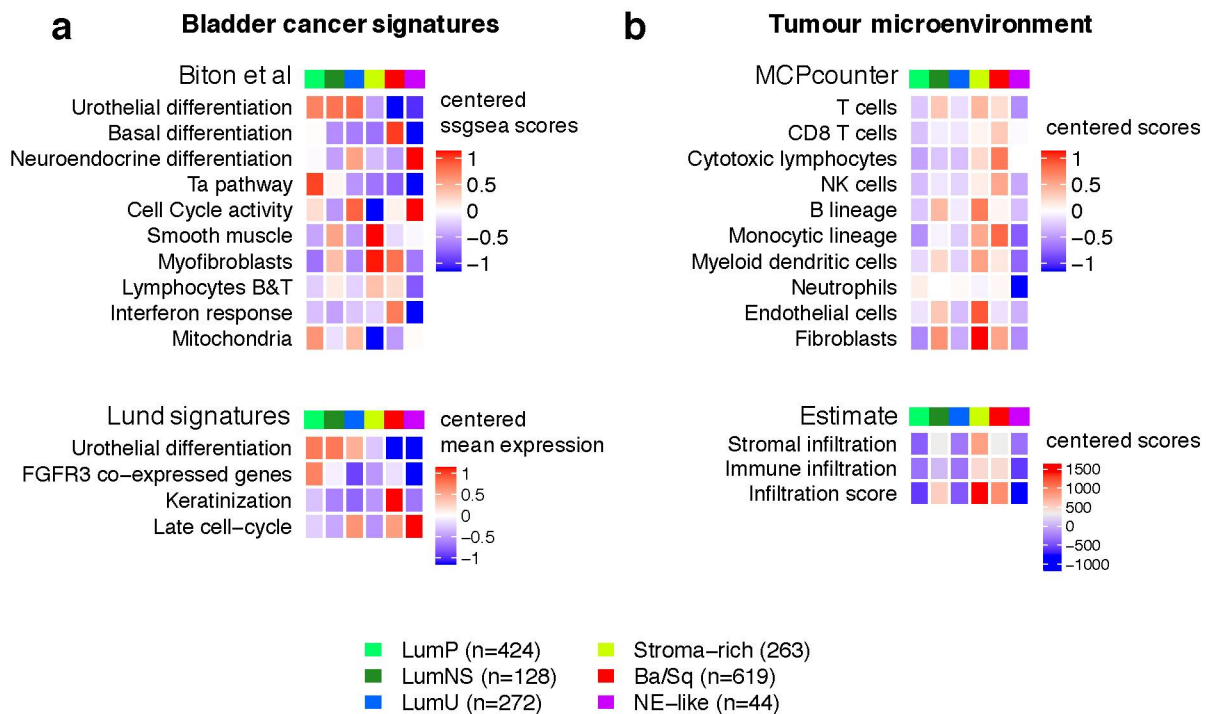
We compared the consensus classes to the 15 TCGA pan-cancer integrative clusters¹⁸ that contained MIBC tumours (Supplementary Figure 2b). We observed enrichments between the Ba/Sq consensus class and the squamous cell carcinoma C27:Pan-SCC pan-cancer cluster ($P = 1.10 \times 10^{-11}$), and between the Stroma-rich class

and the stroma-driven C20: Mixed(Stromal/Immune) pan-cancer cluster ($P < 2.2 \times 10^{-16}$).

Transcriptomic characterization of the six consensus molecular classes

We used mRNA data from all 1750 samples to characterize consensus classes with published molecular gene signatures for bladder cancer pathways and for tumour microenvironment infiltration (Figure 2, Supplementary Table 2).

Figure 2 : Characterization of tumour and stroma signals using published mRNA signatures



Differentiation-associated mRNA signatures were strongly associated with the consensus classes. Tumours from the three luminal classes overexpressed urothelial differentiation signatures ($P < 10^{-16}$), including the PPARG/GATA3/FOXA1-related Lund signature¹⁹. In contrast, Ba/Sq and NE-like tumours respectively overexpressed gene

signatures associated with basal ($P < 10^{-16}$) and neuroendocrine differentiation ($P = 4.2 \times 10^{-16}$).

In addition to their urothelial differentiation status, the three luminal classes exhibited distinct molecular signatures. LumP tumours were characterized by high expression of a non-invasive Ta pathway signature²⁰ ($P < 10^{-16}$) and were strongly associated with FGFR3 transcriptional activity as measured by an FGFR3 co-expressed genes signature⁵ ($P < 10^{-16}$). LumNS tumours displayed elevated stromal infiltration signatures, mainly fibroblastic, as compared to the other luminal tumours ($P < 10^{-16}$). LumU tumours had a high cell cycle activity, and notably overexpressed a late cell cycle signature ($P < 10^{-16}$).

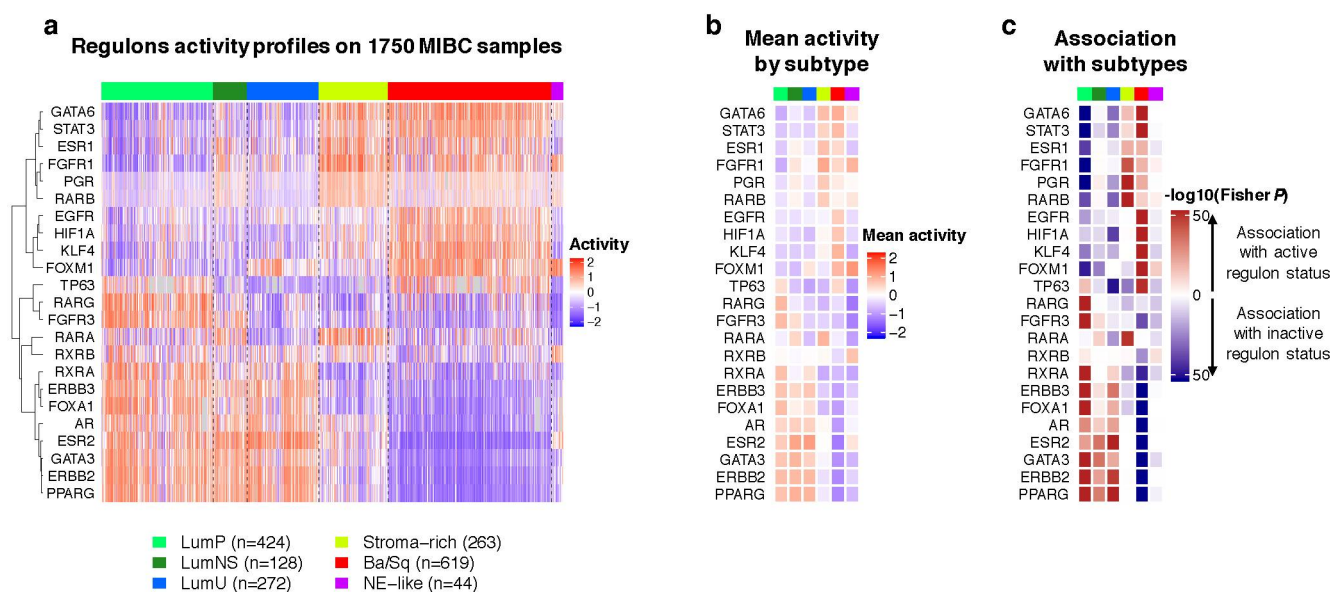
Stroma-rich tumours displayed intermediate and heterogeneous levels of urothelial differentiation. They were mainly characterized by stromal infiltration as summarized by ESTIMATE²¹ stromal scores, with a specific overexpression of smooth muscle and endothelial cell signatures ($P < 10^{-16}$). Fibroblasts and myofibroblasts signatures were also overexpressed within the Stroma-rich tumours ($P < 10^{-16}$).

Immune infiltration was mainly found within Ba/Sq and Stroma-rich tumours, but the two classes were characterized by distinct immune cell populations, as measured by MCPcounter signatures²². Ba/Sq tumours were enriched in cytotoxic lymphocytes and NK cells ($P < 10^{-16}$), whereas Stroma-rich tumours overexpressed T cell and B cell markers ($P < 10^{-16}$). LumNS tumours were the only luminal tumours associated with moderate immune infiltration signals (mainly B and T lymphocytes). We detected no transcriptomic markers of immune infiltration in NE-like tumours.

Analysis of regulatory units (i.e. regulons) for 23 regulator genes previously reported as associated with bladder cancer^{2,23} were consistent with the mRNA signatures assessed (Extended data figure 2). *PPARG* and *GATA3* regulons were

activated within the luminal tumours, which overexpressed strong urothelial differentiation signals. The *FGFR3* regulon was specifically activated within LumP tumours, and Ba/Sq tumours showed the strongest association with the *STAT3* regulon activation, consistent with their expressing a keratinization gene signature. Additionally, the regulon analysis showed an elevated *HIF1A* activity specifically in Ba/Sq tumours, suggesting that this class is associated with a highly hypoxic microenvironment. *EGFR* activity was also specifically associated with Ba/Sq tumours, as previously reported⁸.

Extended data figure 2 : Regulons activity within consensus classes.

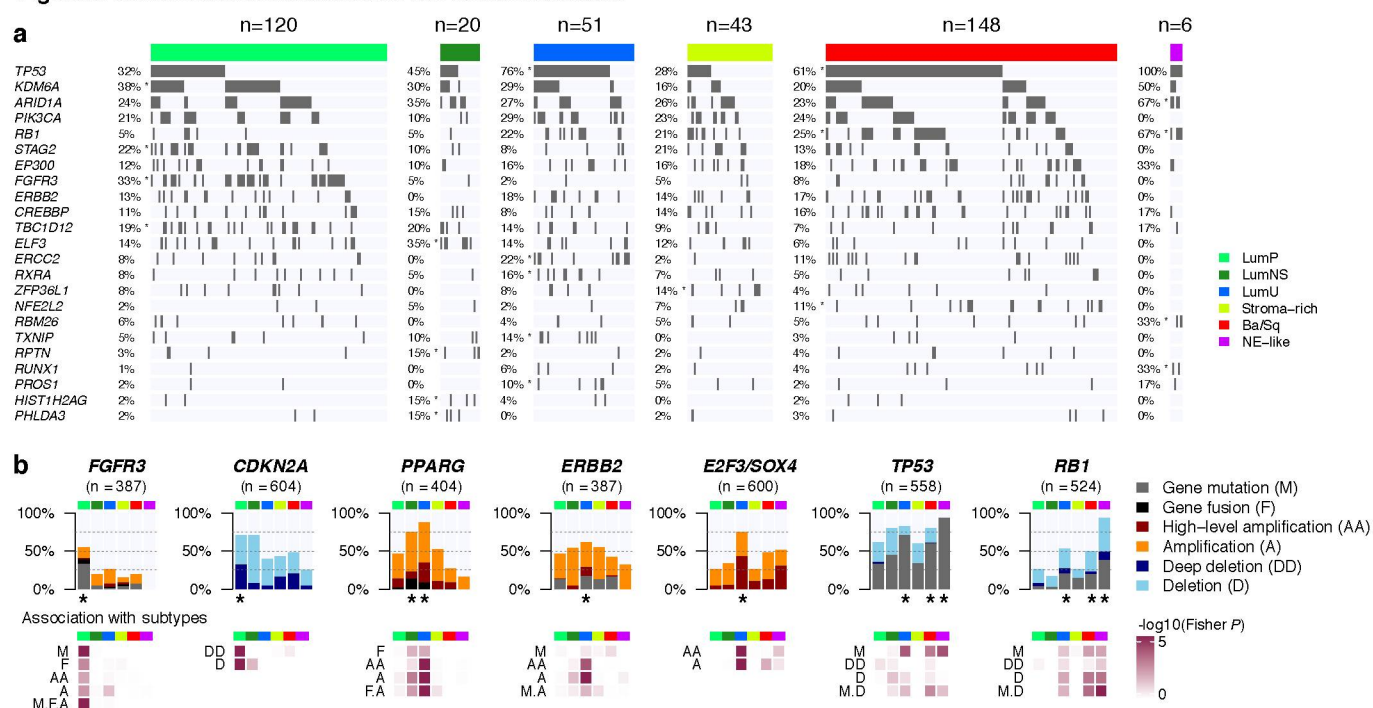


Genomic alterations associated with the consensus molecular classes

We used TCGA exome data to identify class-specific mutations (Figure 3a, Supplementary Table 3) and ran GISTIC2²⁴ on 600 available copy number profiles grouped by consensus class to identify class-specific copy number variations (CNV) (Supplementary Table 4). In addition, we combined all CNV, gene fusion, and gene mutation data from the 18 cohorts to generate comprehensive profiles of genomic

alterations by consensus class, for seven key bladder cancer key genes: *FGFR3*, *CDKN2A*, *PPARG*, *ERBB2*, *E2F3*, *TP53* and *RB1* (Figure 3b).

Figure 3 : Genomic alterations associated with consensus classes

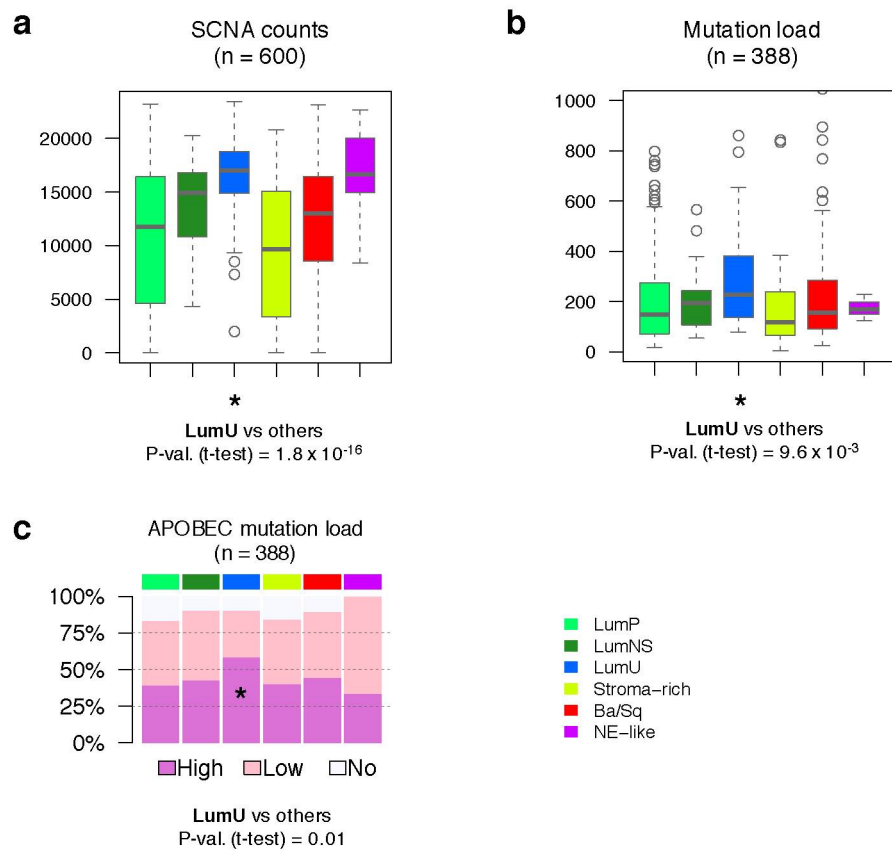


LumP tumours were enriched in *FGFR3* ($P=1.4 \times 10^{-11}$), *KDM6A* ($P=0.002$) and *STAG2* mutations ($P=0.01$). Aggregating data for 643 LumP tumours, the proportion of *FGFR3*-mutated tumours reached 40% ($P=1.6 \times 10^{-23}$). Assembling mutations, fusions, and copy number amplifications, *FGFR3* alterations were enriched in LumP tumours ($P=1.9 \times 10^{-11}$). *CDKN2A* MLPA (Multiplex Ligation-dependent Probe Amplification) and CNV data for 604 tumours revealed 33% of *CDKN2A* homozygous/deep deletions in LumP tumours, corresponding to a strong enrichment as compared to other tumours ($P=3.8 \times 10^{-8}$). These deletions were consistent with the enrichment of LumP tumours within the TCGA pan-cancer iCluster C7: Mixed(Chr9 del) ($P=1.6 \times 10^{-10}$), which is characterized by Chr 9 deletions (Supplementary Figure 2b).

The LumNS class was mainly characterized by an enrichment of mutations in *ELF3* (35%, $P=0.004$), which is an early regulator of normal urothelium, and is activated by PPAR γ ²⁵. *PPARG* was significantly altered as well, with 76% of LumNS tumours harbouring either amplifications or fusions involving this gene ($P=5.7\times 10^{-3}$).

LumU tumours also harboured frequent *PPARG* alterations (89%, $P=1.9\times 10^{-11}$), and high-level amplifications of a 6p22.3 region that contains *E2F3* and *SOX4* (76%, $P=3.0\times 10^{-12}$). Genomic amplifications of *ERBB2* were overrepresented in LumU tumours ($P=4.3\times 10^{-8}$), but no significant association was found between *ERBB2* mutations and any of the consensus classes. In contrast with the other luminal tumours, LumU tumours were associated with *TP53* mutations (76%, $P=3.4\times 10^{-5}$), and with mutations in the core nucleotide-excision repair gene *ERCC2* (22%, $P=0.006$). More generally, LumU was the most genomically altered class (Extended data figure 3), displaying the highest number of copy number alterations ($P=1.8\times 10^{-16}$), the highest somatic mutation load ($P=0.009$), and including more APOBEC-induced mutations than other consensus classes ($P=0.01$). These features of genomic instability and the association with *ERBB2* amplifications were consistent with the enrichment of LumU tumours within the TCGA pan-cancer subtypes C2:BRCA(HER2 amp) (characterized by frequent *ERBB2* amplifications, $P=4.0\times 10^{-5}$) and C13:Mixed(Chr8 del) (enriched in highly aneuploid tumours, $P=3.8\times 10^{-9}$) (Supplementary Figure 2b).

Extended data figure 3: Distributions of SCNA, and total somatic and APOBEC mutation loads across consensus classes



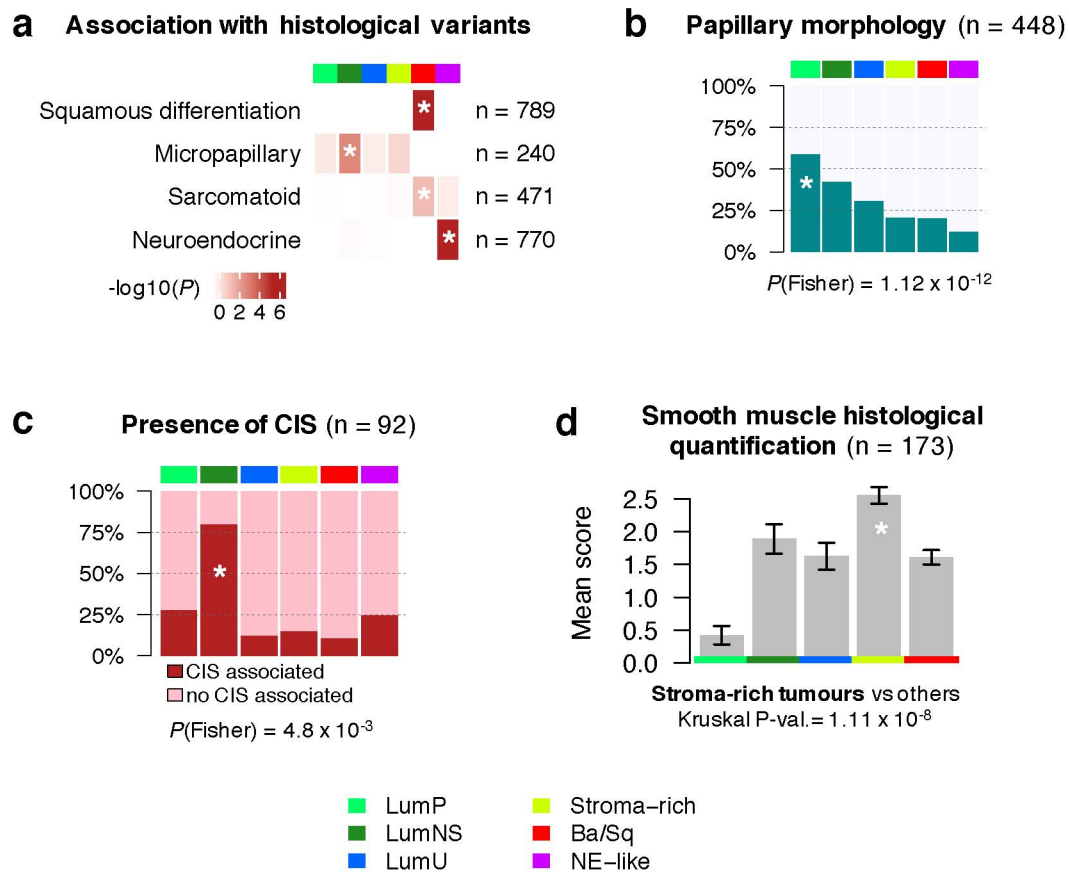
For Ba/Sq tumours, as shown previously²⁶, the most frequently mutated genes were *TP53* ($P=5.8 \times 10^{-4}$), *NFE2L2* ($P=0.002$) and *RB1* ($P=0.002$). Aggregated mutation data revealed that 58% (134/232, $P=0.009$) and 20% (43/224, $P=0.007$) of Ba/Sq tumours contained mutations in *TP53* and *RB1*, respectively. These mutations co-occurred in 14% (32/224) of cases. Ba/Sq tumours were also strongly associated with genomic deletions of a 3p14.2 region, which occurred in 49% of cases ($P = 1.5 \times 10^{-13}$).

Finally, combining all available data on *TP53* and *RB1* genomic alterations, we observed a strong enrichment of *TP53* and *RB1* inactivation in NE-like tumours. *TP53* was ubiquitously mutated in these tumours (94%, $P=9.7 \times 10^{-5}$), and co-occurred with *RB1* inactivation by either mutations or deletions (94%, $P=2.2 \times 10^{-6}$).

Histological patterns associated with the consensus molecular classes

To characterize the consensus molecular classes from a histological perspective, we assembled sample annotations for urothelial histological variants and specific morphologic patterns (Figure 4). As expected, Ba/Sq tumours included 79% of histologically reviewed tumours with squamous differentiation (126/159, $P=3.6 \times 10^{-32}$, Supplementary Figure 3). The Ba/Sq class did however extend beyond this histological subtype, with only 42% (126/303) of Ba/Sq tumours associated with squamous differentiation. Similarly, NE-like tumours were strongly associated with neuroendocrine variant histology, with 72% of histologically reviewed NE-like tumours showing neuroendocrine differentiation (13/18, $P=9.7 \times 10^{-22}$). LumP tumours were enriched with papillary morphology as compared to other consensus classes ($P=1.2 \times 10^{-12}$). This pattern was observed in 59% (82/139) of histologically reviewed LumP tumours, although frequently found in other luminal tumours (42% in LumNS and 31% in LumU tumours). LumNS tumours were enriched in micropapillary variant histology (36%, 9/25, $P=0.001$) and with the presence of carcinoma *in situ* (CIS) lesions (80%, 4/5, $P=0.005$).

Figure 4 : Histopathological associations with consensus classes



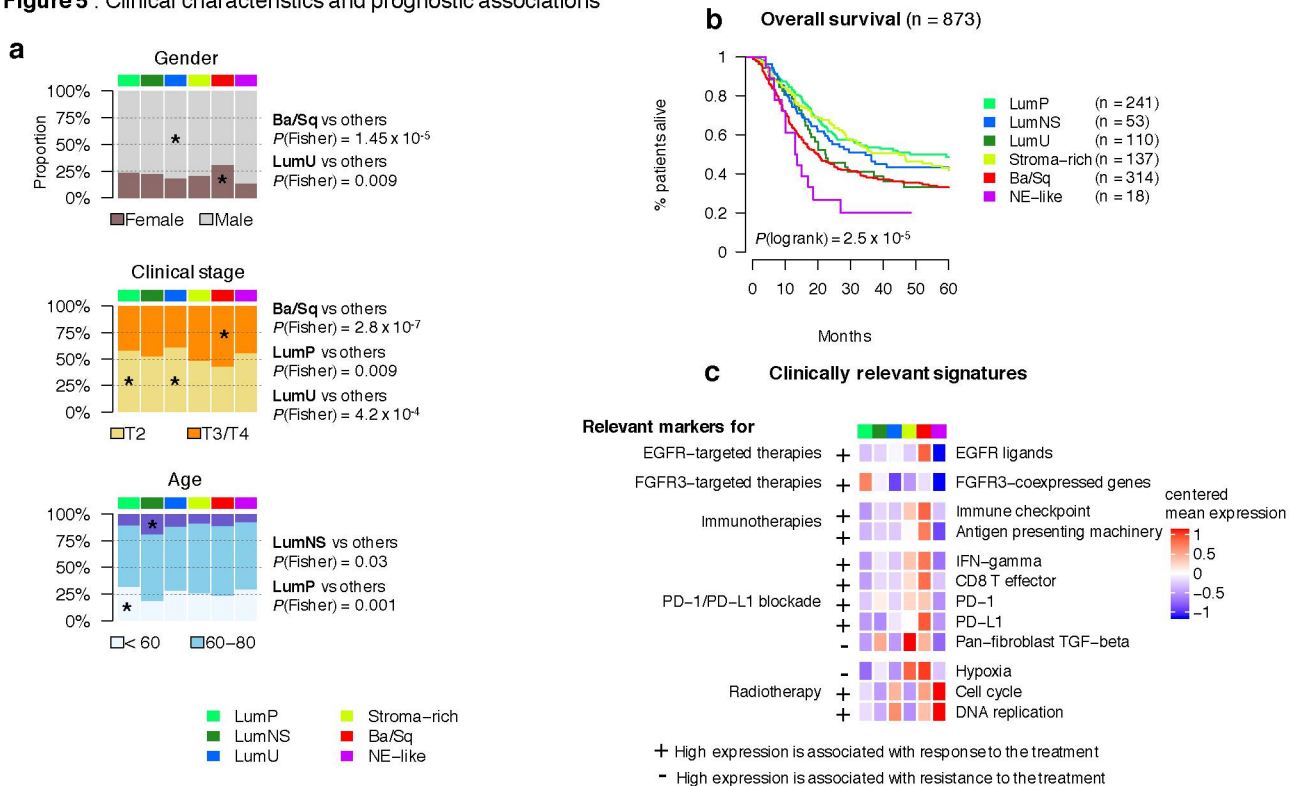
A pathological review of stromal infiltration in TCGA tumour sample slide images confirmed that Stroma-rich tumours were associated with a higher proportion of smooth muscle cells (Kruskal $P=1.1 \times 10^{-8}$), consistent with the strong smooth muscle-related mRNA expression characterizing these tumours.

The consensus molecular classes are associated with distinct clinical characteristics, survival outcomes, and therapeutic opportunities.

We confirmed previously reported associations with gender, stage, and age (Figure 5a), such as the overrepresentation of Ba/Sq tumours in females and in higher clinical stages ($P=1.4 \times 10^{-5}$ and $P=2.8 \times 10^{-7}$ respectively). The LumP and LumU

consensus classes were enriched in T2 vs T3-4 tumours ($P=0.009$ and $P=4.2 \times 10^{-4}$) as compared to other classes. Patients less than 60 years old were overrepresented among LumP tumours ($P=0.001$), whereas the LumNS consensus class was enriched with older patients (> 80 years old; $P=0.03$).

Figure 5 : Clinical characteristics and prognostic associations



Overall survival was strongly associated with the consensus classes (Figure 5b, $P=2.5 \times 10^{-5}$). Patients with LumP tumours had the best prognosis when compared to all consensus classes ($\text{HR}=0.65$, $P=2.1 \times 10^{-4}$, Supplementary Table 5a). The two other luminal classes were associated with poorer prognoses ($\text{HR}_{\text{LumNS}/\text{LumP}}=1.51$, $P=4.7 \times 10^{-2}$; and $\text{HR}_{\text{LumU}/\text{LumP}}=1.32$, $P=0.12$), although the differences were modest or not significant in this setting. Despite the variable differentiation states among samples from the Stroma-rich class, patients with these tumours showed a similar overall survival to that associated with LumP tumours ($\text{HR}_{\text{Stroma-rich}/\text{LumP}}=1.18$, $\text{CI}_{95} = [0.85,$

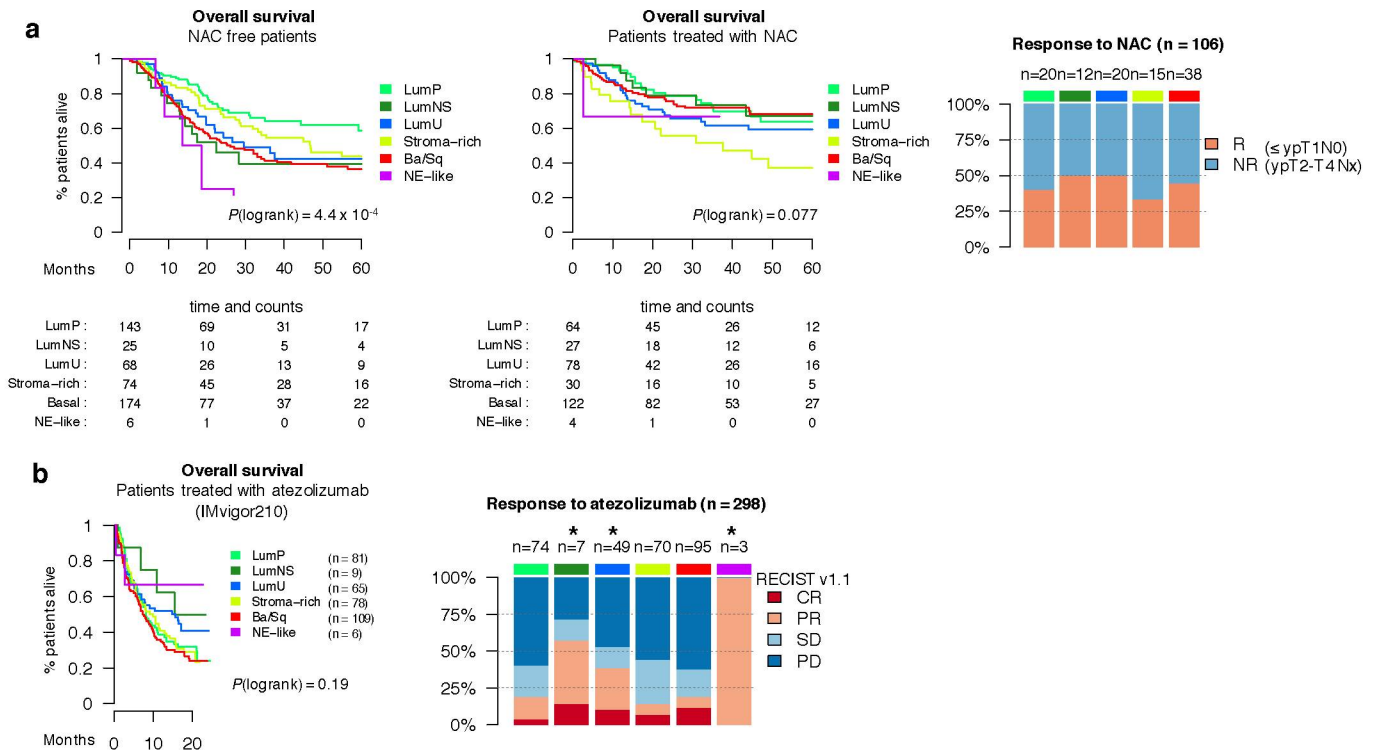
1.63]). Ba/Sq tumours were associated with a poor prognosis ($HR_{Ba/Sq/LumP}=1.8$, $P=5.7 \times 10^{-6}$), consistent with previous studies. Finally, NE-like tumours were associated with the worst prognosis ($HR_{NE-like/LumP}=2.4$, $P=3.3 \times 10^{-3}$). Ba/Sq and NE-like consensus classes remained significantly associated with worse overall survival in a multivariate Cox model that combines consensus classes (with the LumP class as reference), TNM, and patient age (respectively $P=0.002$ and $P=0.05$, Supplementary Table 5b).

We characterized the consensus classes using several clinically relevant mRNA signatures (Figure 5c, Supplementary Table 6). FGFR3 activity signature was strongly and specifically expressed in LumP tumours, suggesting prospects for FGFR3-targeted therapies within this class. Ba/Sq tumours expressed high levels of the EGFR ligands, which may be associated with a sensitivity to EGFR-targeted therapies, as suggested by previously reported *in vitro* and *in vivo* experiments⁸. Ba/Sq tumours also strongly expressed immune checkpoint markers and antigen-presenting machinery genes, suggesting possibilities for immunotherapies within this class. Studies integrating mRNA signatures with response data to anti-PD1/PD-L1 therapies^{17,27} have reported associations of anti-PD1/PD-L1 response with high levels of CD8 T cells, high interferon gamma signals, and low activity of the TGF-beta pathway; however, no consensus class had an expression profile suggesting either response or resistance to anti-PD1/PD-L1 therapies. In contrast, NE-like and LumU tumours both had a profile associated with response to radiotherapy^{28,29}, showing elevated cell cycle activity and low hypoxia signals.

Finally, we performed a consensus class-based retrospective analysis of clinical outcome from patients receiving neoadjuvant chemotherapy^{9,16} (NAC) and patients treated with the anti-PD-L1 atezolizumab¹⁷ (IMvigor210) (Extended data figure 4). Analysis of overall survival and response showed that consensus classes were

associated with distinct responses to the treatments. The results suggested an improved overall survival in the NAC setting for LumNS, LumU and Ba/Sq tumours, and an enrichment in atezolizumab responders in LumNS ($P=0.05$), LumU ($P=0.003$) and NE-like ($P=0.01$) tumours.

Extended data figure 4 : Response to cisplatin-based Neoadjuvant Chemotherapy (NAC) and PD-L1 blockade (atezolizumab)





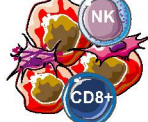



Discussion

The diversity of published MIBC classifications has delayed transferring subtypes into clinical trials or clinical practice. Here, we offer two resources to support work towards such a transfer. First, we analysed the relationships among six different published classification systems, based on 1750 MIBC transcriptomic profiles. We identified six consensus MIBC molecular classes: Basal/Squamous (Ba/Sq) (35%), Luminal Papillary (LumP) (24%), Luminal Unstable (LumU) (15%), Stroma-rich (15%), Luminal Non-Specified (LumNS) (8%), and Neuroendocrine-like (NE-like) (3%). Each

consensus class has distinct differentiation patterns, oncogenic mechanisms, tumour microenvironments, and histological and clinical associations (Figure 6). At this point, NE-like and Ba/Sq classes are the most stably classified, while the three Luminal classes appear to be less clearly defined. Second, we make available an R-based, single-sample classifier that will identify which consensus class a tumour sample's transcriptome corresponds to.

Figure 6 : Summary of main characteristics of the six consensus classes

	24%	8%	15%	15%	35%	3%
	Luminal Papillary	Luminal Non-Specified	Luminal Unstable	Stroma-rich	Basal/Squamous	Neuroendocrine-like
						
Differentiation	Urothelial / Luminal				Basal	Neuroendocrine
Oncogenic mechanisms	FGFR3 ++ CDKN2A -	PPAR-γ ++	PPAR-γ ++ E2F3 +, ERBB2 + Genomic instability		EGFR +	TP53 --, RB1 --, Cellcycle +
Mutations	<i>FGFR3</i> (40%), <i>KDM6A</i> (38%), <i>STAG2</i> (22%)	<i>ELF3</i> (35%)	<i>TP53</i> (76%), <i>ERCC2</i> (22%) TMB +, APOBEC +		<i>TP53</i> (61%), <i>RB1</i> (25%)	<i>TP53</i> (94%) <i>RB1</i> (39%)
Stromal infiltrate		Fibroblasts		Smooth muscle Fibroblasts Myofibroblasts	Fibroblasts Myofibroblasts	
Immune infiltrate				B cells	CD8 T cells NK cells	
Histology	Papillary morphology	Micropapillary variants			Squamous differentiation	Neuroendocrine differentiation
Clinical	T2 stage +	Older patients + (80+)			Women + T3/T4 stage +	
Median overall survival (years)	4	1.8	2.9	3.8	1.2	1

This consensus classification fully concurs with MIBC differentiation-based stratification, revealing tumour classes that are primarily characterized by urothelial differentiation (Luminal classes), basal/squamous differentiation (Ba/Sq) and neuroendocrine differentiation (NE-like). Additional features including genomic alterations, and pathological or clinical characteristics are strongly associated with one or several classes (Figure 6).

LumP tumours are mainly characterized by strong transcriptional activation of FGFR3, involving a genetic mechanism in more than 50% of LumP samples (mutation, fusion, amplification). Papillary morphology is more frequent for these tumours (59%), although found in more than 30% of other luminal tumours. LumP tumours strongly express transcriptomic markers of the Ta pathway, and are consistently associated with the best prognosis among all MIBC tumours. These data suggest that these tumours result from progression of papillary Ta/T1 NMIBC.

The LumNS class included a relatively small number of tumours with characterizing data, precluding using a more descriptive name. Nevertheless, our results point to interesting associations, such as an enrichment in *ELF3* mutations (35%, $n=7/20$, $P=0.004$) and an association with micropapillary morphology (36%, $n=9/25$, $P=0.001$). A weak association with CIS (80%, $n=4/5$, $P=0.005$) is also observed for these tumours. The LumNS tumours are the only luminal tumours expressing stromal and immune signals. Their associated prognosis is the worst of the three luminal classes ($P=0.05$).

LumU tumours display typical features of genomic instability, such as a higher tumour mutation burden that includes more APOBEC-induced mutations, and more copy number alterations. The “Unstable” descriptor for this class refers to the Genomic Unstable tumours from the Lund classification, which are all included within this class. These tumours are particularly enriched in *TP53* (76%) and *ERCC2* (22%) mutations. LumU tumours are associated with high late cell cycle activity, and *ERBB2* activation through mutations or amplification (63%).

Stroma-rich tumours are mainly characterized by high expression of non-tumour cell markers. Smooth muscle cells dominate the infiltration signals associated with these tumours, but endothelial cells and B lymphocytes are also overrepresented. As

assessed by a urothelial differentiation signature¹⁹ and by differentiation-based classification systems, this class contains both luminal and non-luminal tumours (Supplementary Figure 4). However, patients with luminal and non-luminal Stroma-rich tumours have very similar survival, suggesting that although this subgroup is heterogeneous in regards to tumour cell phenotype, stroma could be the main parameter that, given current treatments, drives its clinical features.

Ba/Sq tumours have high expression of basal differentiation markers, and are strongly associated with squamous differentiation. 42% of Ba/Sq tumours have squamous histological variants, and 79% of such variants are observed in Ba/Sq tumours. Although the Ba/Sq tumours are characterized by *KRT14*, *KRT5/6* and lack of *GATA3*, *FOXA1*, and *PPARG*, downregulation of *PPARG* and *GATA3* is not observed in normal basal cells³¹. In this regard the Ba/Sq class is more similar to squamous urothelial metaplasia, consistent with enrichment in the squamous-cell carcinoma-associated C27 pan-cancer iCluster. Ba/Sq tumours express strong fibroblast and myofibroblast infiltration signals, as well as immune infiltration signals from cytotoxic T cells and NK cells. *EGFR* and *STAT3* activation are specific to this class.

The NE-like class includes virtually all (13 of 16, 81%) of tumours with histological neuroendocrine differentiation, and 72% of NE-like tumours have small-cell neuroendocrine variant histology. These tumours show high cell cycle activity, and all have both *TP53* and *RB1* genes inactivated by mutations or deletions. They have the worst prognosis of all MIBC tumours.

We generated the MIBC consensus classification following a procedure similar to that used by Guinney *et al*³² to identify consensus subtypes in colorectal cancer. Given the diverse nature of the six input classification systems that we used to build

the consensus classes (distinct classification methods, strongly varying numbers of classes), we anticipate that the resulting consensus classification captures most of the molecular heterogeneity described, and that it is currently the best consensus solution for MIBC molecular classification.

Except for the Lund sub-stratification, which used IHC, the original subtype classifications analysed in this study were based on transcriptome data, and mainly considered coding transcripts. Considering other types of DNA, RNA or protein data may refine and subdivide the consensus classes further, helping to decipher the diverse biology and heterogeneity of molecular processes within MIBC.

Some bladder tumours show histological and molecular intra-tumour heterogeneity^{33,34}. Our consensus subtyping system addresses inter-tumour heterogeneity and focuses on defining the main molecular subtypes in MIBC. Our transcriptomic classifier will classify tumours according to the dominant class within the tumour sample analysed. However, we recognize that tumour samples may contain multiple subtypes, and we address how such mixtures are likely to interfere with our single-sample classifier by having the classifier report not simply a class label, but also correlation values with all consensus classes. Further studies are required to assess the importance of intra-tumor heterogeneity in prognosis and response to treatment.

The consensus classification suggests possible therapeutic implications. Both the high rate of *FGFR3* mutations and translocations in LumP tumours, and the *FGFR3* activation signature associated with these tumours, suggest that tumours that have an *FGFR3* activation signature may respond to *FGFR* inhibitors, irrespective of the *FGFR3*'s mutation or translocation status. Novel fibroblast growth factor receptor inhibitors have been reported to clinically benefit MIBC patients that harbour mutations

or translocations (about 20% of MIBC patients) and/or overexpression (about 40% of MIBC patients) of the tyrosine kinase receptor *FGFR3*^{35–38}.

Targeting the tumour microenvironment can be an effective option for cancer treatment. Immunotherapy targeting PD1 or PD-L1 immune checkpoints is now included in the standard of care in the US and most of Europe, for patients with locally advanced or metastatic urothelial cancer who relapse after cisplatin-based chemotherapy or are considered cisplatin ineligible, with a 20% objective response rate. A phase 3 clinical trial has demonstrated the efficacy of targeting tumour vasculature in MIBC using an anti-VEGFR2 inhibitor³⁹. The different stromal components within consensus classes, identified by transcriptomic signatures, as well as our analysis of the IMvigor210 data, suggest that our consensus classification should be considered for further clinical studies involving immunotherapy or anti-angiogenic therapy.

Similarities between MIBC consensus classes and other cancer molecular subtypes may also be considered for future basket trials. We showed that such similarities are observed, for instance, between Ba/Sq MIBC tumours, Head and Neck Squamous Cell, Lung Squamous Carcinoma, and Cervical Squamous Cell carcinomas, which were placed together in the C27 PanCanAtlas TCGA cluster. LumU tumours and other *ERBB2*-amplified tumours in breast and stomach cancers were also grouped together in the C2 TCGA PanCanAtlas cluster. More generally, Damrauer *et al* have shown that bladder cancer and breast cancer luminal tumours share molecular similarities⁷. Indeed, in both cancers the luminal subtypes rely on GATA3 and FOXA1, two transcription factors that are necessary for luminal differentiation, and on a nuclear receptor: the estrogen receptor in breast cancer, and PPARG²⁰ in MIBC. Intriguingly, in both cancers there is evidence that the nuclear receptor is involved in differentiation,

while also having protumorigenic effects. Such comparisons across tumour types may help transfer treatment information from tumours bearing similar characteristics into bladder cancer.

We emphasize that we report biological rather than clinical classes, that can be tested for applications in treatment stratification. We offer the classification and the classifier as resources to apply on a single-patient basis in the work required to refine how such classes can best be used clinically. Notably, we propose the consensus classification as a framework for future studies and clinical trials that are intended to identify better predictive markers. Future sub-stratifications may allow defining a system that is more predictive of response to treatments; in such work, the clinical/strategical issue will be to decide the subtype granularity or resolution³⁰ that is appropriate for a specific problem.

Online Methods

Subtyping of MIBC samples according to published MIBC molecular classifications.

The six classification systems were mainly built on transcriptomic data, as follows: Mo *et al*¹³ (Baylor/Tumour differentiation) developed a 18-gene tumour differentiation signature that molecularly define urothelial differentiation, and used this signature to stratify MIBC patients into two groups, namely basal and differentiated.; Damrauer *et al*⁷ (UNC) performed consensus clustering on four aggregated datasets totalling 236 tumours and identified two major clusters, termed luminal-like and basal-like, based on similarities with breast cancer subtypes; Rebouissou *et al*⁸ (CIT-Curie) performed a hierarchical clustering in seven independent datasets including 370 tumours and identified seven meta-clusters (MC) by measuring similarities between

the clusters obtained in each dataset; Choi *et al*⁹ (MDA) identified three subtypes through hierarchical clustering of a 73 tumours dataset, that were named basal, luminal and p53-like relatively to the transcriptomic markers and signatures expressed within each cluster; Marzouka *et al*¹⁰ (Lund) generated gene expression data from 307 tumours and subdivided the cohort into six groups by hierarchical clustering, followed by further sub-stratification into ten levels using immunohistochemistry (IHC); Robertson *et al*² (TCGA) performed a consensus hierarchical clustering of RNA-seq profiles from 412 tumours compiled by TCGA and identified five expression subtypes. Transcriptomic classifiers for Baylor, UNC, MDA, CIT-Curie, Lund, and TCGA classification systems were provided and/or validated by the respective teams. All classifiers were merged into an R package (<https://github.com/cit-bioinfo/BLCAsubtyping>).

We used these classifiers on 18 MIBC mRNA datasets ($N = 1750$ samples) profiled on ten different gene expression platforms (Supplementary Table 1), and assigned each sample to a subtype in each of the six classification systems. 16 datasets were retrieved from public repositories, and two unpublished datasets were shared by L.D. The normalisation method applied on each dataset is detailed in Supplementary Table 1. The six classifiers were applied on each dataset independently.

Network construction and identification of consensus classes

Classification results from the six classifiers were merged for all 18 datasets, and transformed into a binary matrix D of 1750 samples (rows) x 29 classes (columns), where $D(s, c)$ is set to 1 if sample s belongs to class c and 0 otherwise, each row associated with a given sample contains exactly six 1's, reflecting the six class labels

predicted by the six classification systems. For each pair of classes, Cohen's Kappa scores were computed, evaluating the agreement between the two corresponding binary columns of the matrix, i.e. between the 1750 pairs of belong/don't belong class assignments. We could then build a weighted network, with 29 nodes encoding the input molecular subtype, and weighted edges encoding Cohen's Kappa scores. If two subtypes were related by a Cohen's Kappa score < 0 , no edge was built between them, for negative values mean a complete absence of agreement between the two subtypes assignments. To quantify the statistical significance of the remaining edges, we performed hypergeometric tests for overrepresentation of samples classified to one subtype in another. The resulting p-values were adjusted for multiple hypothesis testing using the Benjamini–Hochberg (BH) method, and only edges corresponding to $P < 0.001$ were kept to build the network represented in Figure 1a. Consensus classes were then identified by partitioning this network into clusters using bootstrap iterations and MCL⁴⁰ (Markov cluster algorithm) as described in Guinney *et al*³².

Clustering results were evaluated for MCL inflation factor I ranging from 3 to 15, with 0.3 increments, and using 500 resampling iterations. Each iteration consisted of randomly selecting 80% of samples before constructing the network and running MCL for a given inflation factor. For each inflation factor, we calculated a consensus clustering matrix, defined by the frequency that each pair of input subtypes is partitioned into the same cluster over all iterations. To evaluate clustering robustness, we computed a mean weighted silhouette width (MWSW) for each clustering result as previously described³². The weighted silhouette width extends the silhouette width by giving more weight to subtypes that are more stable within their assigned clusters. Here, we computed a stability score for each input subtype, defined as the average

frequency over all iterations that its within-cluster associations with other subtypes is the same as predicted by MCL on the network generated with all samples. We then used these stability scores as subtypes' weights to compute weighted silhouette width, and considered the mean over all subtypes as a measure of clustering robustness (Supplementary Figure 1a). Clustering generated four- to six-cluster solutions, all of them yielding a mean silhouette width > 0.95 for at least one inflation factor value (Supplementary Figure 1b). The K=4 solution was very robust (MWSW = 0.99) but poorly informative, revealing one cluster of basal subtypes, one cluster of luminal subtypes, one cluster of infiltrated classes, and one cluster of neuroendocrine-associated subtypes. The K=5 solution isolated an additional cluster containing only two subtypes (CIT MC7 and TCGA Luminal subtypes), which was not enough to clearly define a consensus class. The K=6 solution generated robust clusters (MWSW = 0.95), all containing a minimum of 3 subtypes). Heatmaps of consensus matrices for the three solutions illustrate the robustness of the clusters (Supplementary Figure 1b).

Identification of a core set of consensus samples

For each MIBC sample we performed a hypergeometric test for overrepresentation of the sample's assigned input subtypes in the set of input subtypes associated with each consensus class. A sample was assigned to a consensus class if the corresponding overrepresentation test was significant ($P < 0.001$). Using this approach, a core set of 1084 samples were identified to be highly representative of one of the 6 consensus classes and were labelled as consensus samples. We used these consensus samples to build and validate a single-sample mRNA classifier for the consensus classes, then used this classifier to assign consensus labels to all 1750 MIBC samples.

Single-sample transcriptomic classifier construction

We performed feature selection using a training core set of consensus samples from SjödaHL2017 (n=129) and TCGA (n=274) mRNA datasets, both of these sample sets including at least three consensus samples for each consensus class. In each dataset we performed LIMMA moderated t-tests (limma_3.39.1 R package) for each consensus class relative to the others and computed the AUC associated with each gene for the prediction of each class. We summarized the results for each gene common to both datasets (n=17381), using Stouffer's method to aggregate p-values, and computing a mean fold-change for each class comparison. For each class, we selected the genes with Stouffer $P < 0.05$ and $AUC > 0.6$ in at least one of the two datasets, and ordered them according to their mean fold-change. We used these ordered gene lists to generate several lists of varying sizes, by selecting the N top upregulated genes and the N top down-regulated genes in each consensus class, with N varying from 10 to 125. A Pearson nearest-centroid classifier was built on the 129 SjödaHL2017 core samples for each of these gene lists, and its mean balanced accuracy was tested on the independent 681 consensus samples that had not been used for feature selection. The gene list that optimized mean balanced accuracy (97.23%, Supplementary Figure 5) comprised 857 unique genes, and was used to build the final classifier. Six centroids corresponding to the six consensus classes (i.e. the mean mRNA profile of the 857 genes over each consensus class) were computed on the 129 consensus samples from SjödaHL2017 dataset. To classify the 1750 samples into one of the consensus classes, a Pearson correlation was computed between each sample and each centroid. Each sample was then assigned the consensus class whose centroid was the most correlated with the sample profile. If the maximal

correlation for a given sample was less than 0.2, no consensus class label was assigned. This Pearson-based approach does not require to add a pre-processing step to the usual batch normalization of gene expression data, as long as the data are log-transformed, and can therefore be used in a single-sample setting. As shown in Supplementary Figure 5c, the classifier accuracy was similar when using Affymetrix, Illumina, or RNA-seq data. The classifier is publicly available as an R package at <https://github.com/cit-bioinfo/consensusMIBC>.

Comparison with TCGA pan-cancer classifications.

The consensus bladder cancer classification scheme was compared to the TCGA's PanCancerAtlas pan-cancer subtypes¹⁸. We visualized the overlap of classification schemes by calculating the percentage within each MIBC consensus class across the TCGA PanCancer Atlas iCluster classification. We then normalized each row (consensus class) by setting the sum of squares equal to 1. We clustered these data using 1-pearson correlation and used a heatmap for visualisation. To evaluate the significance of the enrichment of consensus classes with certain pan-cancer classifications, we calculated the Chi-Square or Fisher's Exact test p-value from a 2x2 contingency table for the given two classifications of interest. To account for multiple testing, we calculated the Bonferroni p-value threshold for 441 pairwise comparisons to be $P < 0.00011$.

Extraction of bladder cancer gene signatures from Biton *et al*

In their study, Biton *et al* identified and characterized several major bladder cancer signals by an independent component analysis of bladder cancer transcriptome data²⁰. We used ten of these independent components to extract gene sets associated

with the Ta pathway (CIT-13), basal differentiation (CIT-6), cell cycle (CIT-7), urothelial differentiation (CIT-9), smooth muscle (CIT-3), lymphocytes B&T (CIT-8), myofibroblasts (CIT-12), interferon response (CIT-5), neuroendocrine differentiation (CIT-18), and mitochondria (CIT-4). We retrieved the sample contribution vectors associated with each of these components and correlated these values to each gene of the CIT mRNA dataset. Genes that had a Pearson correlation greater than 0.6 (or less than -0.6, depending on the direction of the component association with the biological signal) were selected as representative gene sets for the biological signals associated to the component. The resulting gene sets are given in Supplementary Table 2. For each mRNA dataset included in the study, the R package GSVA⁴¹ (1.30.0) was used to compute single-sample GSEA (Gene Set Enrichment Analysis) scores for the 10 gene sets obtained. The scores were scaled and centered by gene in order to aggregate all datasets. Mean scores were then computed for each consensus class.

Computation of regulon activity scores for 23 regulators

A transcriptional regulatory network for 23 regulators reported as associated with bladder cancer was reconstructed from the TCGA (n=404) MIBC RNA-seq data² using the RTN R package (2.6.0). This regulatory network reconstruction was provided as an RTN TNI-class object, and used to calculate regulon activity scores for 18 cohorts, individually. In each sample in each cohort, for each regulon we used RTN's `tni.gsea2` function to calculate two-tailed GSEA tests²³. This generated regulon activity profiles (RAPs) for each cohort; such a profile shows regulon activities of samples, relative to other samples in the same cohort. Regulons were also assigned discrete status as 'activated', 'neutral' and 'inactivated' in each sample based on their activity.

Statistical analyses

We measured association between consensus classes and categorical variables by Fisher's exact or Chi-square tests. We evaluated differences of continuous variables distributions between consensus classes by Kruskal-Wallis tests, ANOVA or LIMMA moderated t-tests (limma_3.39.1 R package).

We built multivariate Cox models integrating consensus classes and clinical risk factors, stratified on cohort of patients (separate baseline hazard functions were fit for each strata). We used Wald tests to assess survival differences associated with different levels of a given factor included in the Cox models. For each factor level, we computed Hazard Ratios (HR) and 95% Confidence Intervals (CI). We constructed Kaplan-Meier curves to visualize overall survival stratified by consensus class and used log-rank tests to compare the survival of corresponding patient groups.

All statistical and bioinformatics analyses were performed with R software environment (version 3.5.1).

References

1. Knowles, M. A. & Hurst, C. D. Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity. *Nat. Rev. Cancer* **15**, 25–41 (2015).
2. Robertson, A. G. *et al.* Comprehensive Molecular Characterization of Muscle-Invasive Bladder Cancer. *Cell* **171**, 540-556.e25 (2017).
3. Blaveri, E. *et al.* Bladder cancer outcome and subtype classification by gene expression. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **11**, 4044–4055 (2005).
4. Lindgren, D. *et al.* Combined gene expression and genomic profiling define two intrinsic molecular subtypes of urothelial carcinoma and gene signatures for molecular grading and outcome. *Cancer Res.* **70**, 3463–3472 (2010).
5. Sjödaahl, G. *et al.* A molecular taxonomy for urothelial carcinoma. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **18**, 3377–3386 (2012).
6. Volkmer, J.-P. *et al.* Three differentiation states risk-stratify bladder cancer into distinct subtypes. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 2078–2083 (2012).
7. Damrauer, J. S. *et al.* Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 3110–3115 (2014).
8. Rebouissou, S. *et al.* EGFR as a potential therapeutic target for a subset of muscle-invasive bladder cancers presenting a basal-like phenotype. *Sci. Transl. Med.* **6**, 244ra91 (2014).
9. Choi, W. *et al.* Identification of distinct basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy. *Cancer Cell* **25**, 152–165 (2014).
10. Marzouka, N. *et al.* A validation and extended description of the Lund taxonomy for urothelial carcinoma using the TCGA cohort. *Sci. Rep.* **8**, 3737 (2018).

11. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
12. Sjödaahl, G., Eriksson, P., Liedberg, F. & Höglund, M. Molecular classification of urothelial carcinoma: global mRNA classification versus tumour-cell phenotype classification. *J. Pathol.* **242**, 113–125 (2017).
13. Mo, Q. *et al.* Prognostic Power of a Tumor Differentiation Gene Signature for Bladder Urothelial Carcinomas. *J. Natl. Cancer Inst.* (2018).
doi:10.1093/jnci/djx243
14. Hedegaard, J. *et al.* Comprehensive Transcriptional Analysis of Early-Stage Urothelial Carcinoma. *Cancer Cell* **30**, 27–42 (2016).
15. Rosenberg, J. E. *et al.* Atezolizumab in patients with locally advanced and metastatic urothelial carcinoma who have progressed following treatment with platinum-based chemotherapy: a single-arm, multicentre, phase 2 trial. *The Lancet* **387**, 1909–1920 (2016).
16. Seiler, R. *et al.* Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy. *Eur. Urol.* **72**, 544–554 (2017).
17. Mariathasan, S. *et al.* TGF β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature* **554**, 544–548 (2018).
18. Hoadley, K. A. *et al.* Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell* **173**, 291-304.e6 (2018).
19. Eriksson, P. *et al.* Molecular subtypes of urothelial carcinoma are defined by specific gene regulatory systems. *BMC Med. Genomics* **8**, 25 (2015).

20. Biton, A. *et al.* Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.* **9**, 1235–1245 (2014).
21. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
22. Becht, E. *et al.* Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol.* **17**, 218 (2016).
23. Castro, M. A. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat. Genet.* **48**, 12–21 (2016).
24. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
25. Böck, M. *et al.* Identification of ELF3 as an early transcriptional regulator of human urothelium. *Dev. Biol.* **386**, 321–330 (2014).
26. Choi, W. *et al.* Genetic Alterations in the Molecular Subtypes of Bladder Cancer: Illustration in the Cancer Genome Atlas Dataset. *Eur. Urol.* **72**, 354–365 (2017).
27. Ayers, M. *et al.* IFN- γ -related mRNA profile predicts clinical response to PD-1 blockade. *J. Clin. Invest.* **127**, 2930–2940 (2017).
28. Pawlik, T. M. & Keyomarsi, K. Role of cell cycle in mediating sensitivity to radiotherapy. *Int. J. Radiat. Oncol. Biol. Phys.* **59**, 928–942 (2004).
29. Horsman, M. R. & Overgaard, J. The impact of hypoxia and its modification of the outcome of radiotherapy. *J. Radiat. Res. (Tokyo)* **57**, i90–i98 (2016).

30. Aine, M., Eriksson, P., Liedberg, F., Höglund, M. & Sjö Dahl, G. On Molecular Classification of Bladder Cancer: Out of One, Many. *Eur. Urol.* **68**, 921–923 (2015).
31. Fishwick, C. *et al.* Heterarchy of transcription factors driving basal and luminal cell phenotypes in human urothelium. *Cell Death Differ.* **24**, 809–818 (2017).
32. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **advance online publication**, (2015).
33. Warrick, J. I. *et al.* Intratumoral Heterogeneity of Bladder Cancer by Molecular Subtypes and Histologic Variants. *Eur. Urol.* **0**, (2018).
34. Thomsen, M. B. H. *et al.* Comprehensive multiregional analysis of molecular heterogeneity in bladder cancer. *Sci. Rep.* **7**, 11702 (2017).
35. Taberero, J. *et al.* Phase I Dose-Escalation Study of JNJ-42756493, an Oral Pan-Fibroblast Growth Factor Receptor Inhibitor, in Patients With Advanced Solid Tumors. *J. Clin. Oncol.* **33**, 3401–3408 (2015).
36. Nogova, L. *et al.* Evaluation of BGJ398, a Fibroblast Growth Factor Receptor 1-3 Kinase Inhibitor, in Patients With Advanced Solid Tumors Harboring Genetic Alterations in Fibroblast Growth Factor Receptors: Results of a Global Phase I, Dose-Escalation and Dose-Expansion Study. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **35**, 157–165 (2017).
37. Schuler, M. *et al.* 859P Anti-tumor activity of the pan-FGFR inhibitor rogaratinib in patients with advanced urothelial carcinomas selected based on tumor FGFR mRNA expression levels. *Ann. Oncol.* **28**, (2017).
38. Pal, S. K. *et al.* Efficacy of BGJ398, a fibroblast growth factor receptor 1-3 inhibitor, in patients with previously treated advanced urothelial carcinoma with

FGFR3 alterations. *Cancer Discov.* CD-18-0229 (2018). doi:10.1158/2159-8290.CD-18-0229

39. Petrylak, D. P. *et al.* Ramucirumab plus docetaxel versus placebo plus docetaxel in patients with locally advanced or metastatic urothelial carcinoma after platinum-based therapy (RANGE): a randomised, double-blind, phase 3 trial. *The Lancet* **390**, 2266–2277 (2017).
40. Van Dongen, S. Graph Clustering Via a Discrete Uncoupling Process. *SIAM J. Matrix Anal. Appl.* **30**, 121–141 (2008).
41. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).

Figure legends

Figure 1: The six consensus classes and their relation to input molecular subtypes. (a) Clustered network by MCL clustering. The 6-consensus classes solution obtained with MCL clustering on the Cohen's Kappa-weighted network is represented by the 6 cliques surrounded by black dotted rectangles. The circles inside each clique symbolize the input subtypes associated with each consensus class and are coloured according to their matching classification system. Circle size is proportional to the number of samples assigned to the subtype. Edge width between subtypes is proportional to the Cohen's Kappa score, which assess the level of agreement between two classification schemes. (b) Input subtypes repartitioned among each consensus class. Consensus classes were predicted on 1750 MIBC samples using the single-sample classifier described in Methods. Here, the samples are grouped by their predicted consensus class label: LumP, LumN, LumU, Stroma-rich, Ba/Sq and Neuroendocrine (NE)-like. For each consensus class, a barplot shows the proportion of samples assigned in each input subtype of each input classification system.

Figure 2: Characterization of tumour and stroma signals using published mRNA signatures. (a) The 1750 mRNA expression profiles were used to compute (above, Biton) mean enrichment scores for specific gene signatures in each consensus class (based on a single-sample GSEA approach), or (below, Lund) mean expression of gene sets. Bladder cancer gene signatures include those related to the ICA components described in Biton *et al*²⁰ (see Methods), as well as other bladder cancer-specific signatures retrieved from the literature: urothelial differentiation,

keratinization and late cell-cycle signatures from Eriksson *et al*⁴², and an *FGFR3* co-expressed signature from Sjödaahl *et al*⁵. (Supplementary Table 2) (b) Tumour microenvironment characterization includes (above) an estimate of microenvironment immune and stromal cell subpopulations using MCPcounter²² and (below) a more global measure of stromal and immune infiltrates by ESTIMATE²¹.

Figure 3: Genomic alterations associated with consensus classes. (a) We used the available exome data from 388 TCGA samples to study the association between consensus classes and specific gene mutations. The panel displays the 23 genes with significant mutations (MutSig $P < 0.001$) that were either found in at least 10% of all tumours, or significantly overrepresented within one of the consensus classes (Fisher $P < 0.05$ and frequency within a consensus class $> 10\%$). Gene mutations that were significantly enriched in one consensus class are marked by an asterisk. (b) Combined genomic alterations associated with seven bladder cancer-associated genes and statistical association with consensus classes. Upper panels: Main alteration types after aggregating CNV profiles from CIT (n=87), Iyer (n=58), Sjödaahl (n=29), Stransky (n=22), and TCGA (n=404) data; exome profiles (n=388) and *FGFR3* and *PPARG* fusion data (n=404) from TCGA data; *CDKN2A* and *RB1* MLPA data from CIT (n=86; n=85) and Stransky (n=16; n=13) data; *FGFR3* mutation data from MDA (n=66), CIT (n=87), Iyer (n=39), Sjödaahl (n=28), and Stransky (n=35) data; TP53 mutation data from MDA (n=66), CIT (n=87), Iyer (n=39), Sjödaahl (n=28), and Stransky (n=19) data; and *RB1* mutation data from MDA (n=66), CIT (n=85), Iyer (n=39) and Stransky (n=13) data. Lower panels: Associations between each consensus class, each type of gene alteration, and the combined alterations were

evaluated by Fisher's exact tests. Consensus classes significantly enriched with alterations of these candidate genes are marked with a black asterisk.

Figure 4: Histopathological associations with consensus classes. (a)

Histological variant overrepresentation within each consensus class. One-sided Fisher exact tests were performed for each class and histological pattern.

Pathological review of histological variants was available for several cohorts: squamous differentiation was evaluated in CIT (n=75), MDA (n=46), Sjødahl2012 (n=23), Sjødahl2017 (n=239) and TCGA (n=406) cohorts; neuroendocrine variants were reviewed in CIT (n=75), MDA (n=46), Sjødahl2017 (n=243), and TCGA (n=406) cohorts; micropapillary variants were reviewed in CIT (n=75), MDA (n=46) and TCGA cohorts (n=118 FFPE tumour slides from TCGA were reviewed by Y.A. and J.F. for this study). Results are displayed on the heatmap as $-\log_{10}(\text{Fisher's } P)$. (b)

The proportion of samples with carcinoma *in situ* (CIS) associated within each consensus class, for 84 tumours from CIT cohort and 8 tumours from Dyrskjøet cohort.

(c) The presence/absence of a papillary morphology, for 401 tumours from TCGA cohort and 47 tumours from CIT cohort. (d) Smooth muscle infiltration from images for 174 tumour slides from the TCGA cohort: 73 LumP, 18 LumNS, 16 LumU, 20 Stroma-rich and 46 Ba/Sq tumour samples. Each sample was assigned a semi-quantitative score ranging from 0 to 3 (0 = absent, 1 = low, 2 = moderate, 3 = high) to quantify the presence of large smooth muscle bundles. The barplot shows means and standard deviations.

Figure 5: Clinical characteristics and prognostic associations. (a) Association of consensus classes with gender (n=1554), clinical stage (n=1641), and age category

(n=1378). **(b)** 5-year overall survival stratified by consensus class. Kaplan-Meier curves were generated from 873 patients with available follow-up data. Patients who were marked as having received neoadjuvant chemotherapy were excluded from the survival analysis. **(c)** The 1750 mRNA expression profiles were used to compute per-class mean expression of gene sets that are clinically relevant for response to therapies (Supplementary Table 6). Gene sets are annotated with a plus (respectively minus) sign if high expression of the genes is associated with response (respectively resistance) to the category of therapies indicated on the left.

Figure 6: Summary of main characteristics of the consensus classes. Top to bottom: Proportion of consensus classes in the n=1750 tumour samples. Consensus classes names. Cellular schematics for tumour cells and their microenvironment (Immune cells, fibroblasts, and smooth muscle cells). Differentiation-based color scale showing the differentiation status associated with consensus classes, including a Luminal-to-basal gradient, and neuroendocrine differentiation status. Table of dominant characteristics: oncogenic mechanisms, mutations, stromal infiltrate, immune infiltrate, histological observations, clinical characteristics, and median overall survival.

Extended data figure legends

Extended data figure 1: Analytical workflow. We used mRNA classifiers provided by 6 teams involved in previously published classification systems to subtype 1750 mRNA profiles from 18 independent MIBC cohorts. A total number of 29 subtypes were considered when summing all classification systems. Using the subtyping results, we could build a 1750 x 29 binary matrix D where a sample s was given a value of 1 if assigned to the subtype m , and 0 otherwise. The matrix D was used to build a network interconnecting the 29 distinct subtypes. Edges between two subtypes were weighted using a Cohen's Kappa metric. We performed MCL clustering⁴⁰ on this network with 500 bootstrap iterations for several values of inflation factors and used stability scores as weights to calculate weighted silhouette width for each resulting cluster. We then used the mean weighted silhouette width as a performance measure to select an inflation factor yielding a robust consensus clustering solution. An optimal consensus solution was reached for 6 consensus classes, and this solution also defined a set of 1084 'core' consensus samples with subtype labels that were highly concordant among the consensus classes ($P < 0.001$, hypergeometric test). We used these 1084 core samples to build a nearest-centroid, single-sample classifier based on Pearson's correlation coefficient, then used the resulting classifier to predict consensus classes on all 1750 MIBC samples. We further characterized the consensus classes using-molecular, histological and clinical data.

Extended data figure 2: Regulon activity within consensus classes. We computed regulon activity profiles (RAPs) as described in Robertson *et al*², for 23

bladder cancer regulators. **(a)** Heatmap of RAPs at the sample level. RAPs were computed on each of the 18 datasets independently and pooled for the heatmap visualization. **(b)** Summary showing the mean RAP for each consensus class. **(c)** Association of a regulon's active or inactive status with each class, indicated by $-\log_{10}(\text{Fisher } P \text{ value})$. Fisher exact tests were done using RAPs that had been discretized by status (1 for active regulon status, 0 for neutral status, -1 for inactive regulon status).

Extended data figure 3: Distributions of SCNA, and total somatic and APOBEC mutation loads across consensus classes **(a)** Distribution of Somatic Copy Number Alteration (SCNA) counts across consensus classes. SCNA counts are defined as the number of genes with copy number changes, as estimated by GISTIC2²⁴ over 600 MIBC CNV profiles from datasets from CIT (n=87), Iyer (n=58), Sjö Dahl2012 (n =29), Stransky (n=22) and TCGA (n=404). **(b)** Distribution of nonsynonymous somatic mutation events across consensus classes. **(c)** Enrichment of APOBEC-induced mutation within consensus classes. The minimum estimate of the number of APOBEC-induced mutations was computed for 388 samples of TCGA MIBC cohort and discretized into categorical values : "No" : estimate = 0; "Low": estimate \leq median of non-zero values (median was 61.5); "High": estimate $>$ median of non-zero values.

Extended data figure 4: Response to neoadjuvant chemotherapy and PD-L1 blockade.

To further explore the association of the consensus classification with therapeutic response, we analysed overall survival and response data from patients who had

received neoadjuvant chemotherapy^{9,16} (NAC) and patients treated with the anti PD-L1 atezolizumab¹⁷ (IMvigor210). The pre-treatment tumour samples from these patients were classified according to the consensus molecular classification. To better evaluate the effect of NAC on overall survival we selected a set of NAC-free patients and compared the class-associated overall survival of these patients with survival of patients receiving NAC. **(a)** Overall survival and response data to neoadjuvant chemotherapy (NAC). For the analysis of overall survival, NAC-free patients were selected from MDA (n=46), Sjødahl (n=51) and TCGA (n=394) cohorts, patients treated with NAC from Seiler (n=273), MDA MVAC (n=22, GSE70691), and MDA DDMVAC (n=38, GSE69795) cohorts.

Pathological response to NAC was obtained from MDA MVAC (n=23), MDA DDMVAC (n=34) and Seiler (n=43) cohorts. **(b)** Overall survival and response to PD-L1 blockade (atezolizumab), from IMvigor210 trial (Mariathasan et al). Consensus classes were predicted for all MIBC samples included in IMvigor210 dataset using the single-sample classifier. Consensus classes associated (Fisher $P < 0.05$) with positive response to atezolizumab, i.e. complete (CR) or partial responders (PR), are indicated by a black asterisk.

Bladder Cancer Molecular Taxonomy Group and affiliations

Mattias Aine, Division of Molecular Hematology, Department of Laboratory Medicine, Faculty of Medicine, Lund University, Lund, Sweden

Hikmat Al-Ahmadie, Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA

Yves Allory, Department of Pathology, Institut Curie Hospital Group, Paris, France

Joaquim Bellmunt, Bladder Cancer Center, Dana-Farber/Brigham and Women's Cancer Center, Harvard Medical School, Boston, MA, 02215, USA

Isabelle Bernard-Pierrot, Oncologie Moléculaire, CNRS UMR 144, Institut Curie, Paris, France

Peter C. Black, Department of Urologic Sciences, University of British Columbia, Vancouver, British Columbia, Canada

Mauro A. A. Castro, Bioinformatics and Systems Biology Laboratory, Federal University of Paraná, Polytechnic Center, Curitiba, Brazil

Keith S. Chan, Molecular & Cellular Biology/Scott Department of Urology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Woonyoung Choi, Johns Hopkins Greenberg Bladder Cancer Institute and Brady Urological Institute, Johns Hopkins University, Baltimore, MD, USA

Bogdan Czerniak, Department of Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA.

Colin P. Dinney, Department of Urology and Department of Cancer Biology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Lars Dyrskjöt, Department of Molecular Medicine, Aarhus University Hospital, Aarhus 8200, Denmark

Pontus Eriksson, Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden

Jacqueline Fontugne, Department of Pathology, Institut Curie Hospital Group, Paris, France

Ewan A. Gibb, GenomeDx Biosciences Inc., Vancouver, BC Canada

Clarice S. Groeneveld, Bioinformatics and Systems Biology Laboratory, Federal University of Paraná, Polytechnic Center, Curitiba, Brazil

Arndt Hartmann, Institute of Pathology, University Erlangen-Nürnberg, Krankenhausstr 8-10, Erlangen, Germany

Katherine A. Hoadley, Department of Genetics, Department of Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

Mattias Höglund, Division of Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden

Aurélié Kamoun, Cartes d'Identité des Tumeurs Program, Ligue Nationale Contre le Cancer, 75013 Paris, France

Jaegil Kim, Broad Institute of MIT and Harvard, Cambridge, MA, USA.

William Y. Kim, Department of Genetics, Department of Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

David Kwiatkowski, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA.

Thierry Leuret, Department of Urology, University of Versailles-Saint-Quentin-en-Yvelines, Foch Hospital, Suresnes, France.

Seth P. Lerner, Scott Department of Urology, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, TX, USA

Fredrik Liedberg, Department of Translational Medicine, Lund University, Skåne University Hospital, Malmö, Sweden

Núria Malats, Genetic and Molecular Epidemiology Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

David J. McConkey, Johns Hopkins Greenberg Bladder Cancer Institute and Brady Urological Institute, Johns Hopkins University, Baltimore, MD, USA

Qianxing Mo, Department of Medicine, Baylor College of Medicine, Houston, TX, USA

Thomas Powles, Barts Cancer Institute ECMC, Barts Health and the Royal Free NHS Trust, Queen Mary University of London, London, UK

François Radvanyi, Oncologie Moléculaire, CNRS UMR 144, Institut Curie, Paris, France

Francisco X. Real, Epithelial Carcinogenesis Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

Aurélien de Reyniès, Cartes d'Identité des Tumeurs Program, Ligue Nationale Contre le Cancer, 75013 Paris, France

A. Gordon Robertson, Canada's Michael Smith Genome Sciences Center, BC Cancer Agency, Vancouver, Canada

Arlene Siefker-Radtke, Department of Genitourinary Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Nanor Sirab, Department of Pathology, Institut Curie Hospital Group, Paris, France

Roland Seiler, Department of Urology, Bern University Hospital, Switzerland

Gottfrid Sjö Dahl, Division of Urological Research, Department of Translational Medicine, Lund University, Skåne University Hospital Malmö, Sweden

Ann Taber, Department of Molecular Medicine, Aarhus University Hospital, Aarhus 8200, Denmark

John Weinstein, Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA;

Alexandre Zlotta, Department of Surgery, Division of Urology, University of Toronto, Mount Sinai Hospital and University Health Network, Toronto, ON, Canada