Title

Integrative pathway enrichment analysis of multivariate omics data.

Authors

Paczkowska, Marta

Barenboim, Jonathan

Sintupisut, Nardnisa

et al.

Integrative pathway enrichment analysis of multivariate omics data

Marta Paczkowska[1,136], Jonathan Barenboim[1,136], Nardnisa Sintupisut[1], Natalie S. Fox[1,2], Helen Zhu[1,2], Diala Abd-Rabbo[1], Miles W. Mee[1], Paul C. Boutros[1,2,3,4,5,6,7,8], PCAWG Drivers and Functional Interpretation Working Group, Jüri Reimand[1,2]* & PCAWG Consortium

Multi-omics datasets represent distinct aspects of the central dogma of molecular biology. Such high-dimensional molecular profiles pose challenges to data interpretation and hypothesis generation. ActivePathways is an integrative method that discovers significantly enriched pathways across multiple datasets using statistical data fusion, rationalizes contributing evidence and highlights associated genes. As part of the ICGC/TCGA Pan-Cancer Analysis of Whole Genomes (PCAWG) Consortium, which aggregated whole genome sequencing data from 2658 cancers across 38 tumor types, we integrated genes with coding and non-coding mutations and revealed frequently mutated pathways and additional cancer genes with infrequent mutations. We also analyzed prognostic molecular pathways by integrating genomic and transcriptomic features of 1780 breast cancers and highlighted associations with immune response and anti-apoptotic signaling. Integration of ChIP-seq and RNA-seq data for master regulators of the Hippo pathway across normal human tissues identified processes of tissue regeneration and stem cell regulation. ActivePathways is a versatile method that improves systems-level understanding of cellular organization in health and disease through integration of multiple molecular datasets and pathway annotations.

[1] Computational Biology Program, Ontario Institute for Cancer Research, 661 University Ave Suite 510, Toronto, ON M5G 0A3, Canada. [2] Department of Medical Biophysics, University of Toronto, 101 College Street Suite 15-701, Toronto, ON M5G 1L7, Canada. [3] Department of Pharmacology & Toxicology, University of Toronto, 1 King's College Circle Room 4207, Toronto, ON M5S 1A8, Canada. [4] Department of Human Genetics, University of California Los Angeles, 10833 Le Conte Avenue, Los Angeles, CA 90095, USA. [5] Department of Urology, University of California Los Angeles, 200 Medical Plaza Driveway #140, Los Angeles, CA 90024, USA. [6] Institute of Precision Health, University of California Los Angeles, 10833 Le Conte Avenue, Los Angeles, CA 90024, USA. [7] Broad Stem Cell Research Centre, University of California Los Angeles, 615 Charles E Young Drive S, Los Angeles, CA 90095, USA. [8] Jonsson Comprehensive Cancer Centre, University of California Los Angeles, 10833 Le Conte Avenue, Los Angeles, CA 90024, USA. [136]These authors contributed equally: Marta Paczkowska, Jonathan Barenboim. PCAWG Drivers and Functional Interpretation Working Group authors and their affiliations appears at the end of the paper. PCAWG Consortium members and their affiliations appear in the Supplementary Information *email: Juri.Reimand@utoronto.ca

Pathway enrichment analysis is an essential step for interpreting high-throughput (omics) data that uses current knowledge of genes and biological processes. A common application determines statistical enrichment of molecular pathways, biological processes and other functional annotations in long lists of candidate genes[1]. Genomic, transcriptomic, proteomic and epigenomic experiments emphasize complementary aspects of underlying biology and are best analyzed integratively, as is now routinely done in large-scale projects such as The Cancer Genome Atlas (TCGA)[2], Clinical Proteome Tumor Analysis Consortium (CPTAC), International Cancer Genome Consortium (ICGC)[3,4], Genotype-Tissue Expression (GTEx)[5], and others[6]. Thus, simultaneous analysis of multiple candidate gene lists for characteristic pathways is increasingly needed.

Numerous approaches are available for interpreting single gene lists. For example, the GSEA algorithm can detect upregulated and downregulated pathways in gene expression datasets[7]. Web-based methods such as Panther[8], ToppCluster[9], and g:Profiler[10] detect significantly enriched pathways amongst ranked or unranked gene lists and are generally applicable to genes and proteins from various analyses. Some approaches allow analysis of multiple input gene lists however these primarily rely on visualization rather than data integration to evaluate the contribution of distinct gene lists towards each detected pathway[9,10]. Finally, no methods are available for unified pathway analysis of coding and non-coding mutations from whole-genome sequencing (WGS) data, or integrating these with other types of DNA aberrations such as copy number changes and balanced genomic rearrangements.

Cancer genomes are characterized by multiple classes of mutations, including single nucleotide variants (SNVs), small insertions–deletions (indels), copy number alterations, and translocations. These affect a small number of frequently mutated pan-cancer driver genes such as *TP53*, less-frequent and tissue-specific genes such as *SPOP* in prostate cancer, and numerous infrequently mutated genes. The majority of currently known driver mutations of SNVs and indels affect protein-coding sequence[11] and only few high-confidence non-coding drivers have been found, such as the mutation hotspots in the *TERT* promoter[12]. Discovery of coding and non-coding driver mutations is a major goal of large cancer whole genome sequencing efforts. The PCAWG Consortium aggregated whole genome sequencing data from 2658 cancers across 38 tumor types generated by the ICGC and TCGA projects. These sequencing data were re-analysed with standardized, high-accuracy pipelines to align to the human genome (reference build hs37d5) and identify germline variants and somatically acquired mutations, as described in the PCAWG marker paper[4]. A consensus analysis of variant calls in PCAWG tumors generated a high-confidence catalog of driver mutations in protein-coding driver genes (CDS) and non-coding regions of 5′ and 3′ untranslated elements (UTRs), promoters and enhancers[13]. A consensus pathway and network analysis of PCAWG driver mutations used knowledge of molecular pathways and gene interaction networks as priors to further discover infrequent candidate driver variants including those in the non-coding genome[14].

Here we report the development of the ActivePathways method that uses data fusion techniques to address the challenge of integrative pathway analysis of multi-omics data. It detects significantly enriched pathways across multiple datasets, including those pathways that are not apparent in any individual dataset. We present several analyses to demonstrate this method. First, we integrate cancer driver genes with coding and non-coding mutations predicted using the PCAWG dataset[13] and reveal numerous processes and additional genes with frequent coding and non-coding mutations. Second, we integrate patient clinical information with transcriptomic and copy number alterations in breast cancers of the METABRIC project[15] to discover prognostic pathways and processes in breast cancer subtypes. Third, we integrate transcriptomic data of normal tissues of the GTEx project[5] with ChIP-seq data to infer gene regulatory networks and biological processes downstream of the Hippo pathway of tissue growth control and regeneration. Thus ActivePathways is a versatile method for combining diverse multiomics datasets.

Results

Multi-omics pathway enrichment analysis with ActivePathways. ActivePathways is a simple three-step method that extends our earlier work[10] (Fig. 1). It requires two input datasets. The first input is a table of *P*-values with genes listed in rows and evidence from distinct datasets listed in columns. The columns can include *P*-values of differential gene expression, gene essentiality, mutation or copy number alteration burden and many others that are derived using platform-specific quantification methods. The second input is a collection of gene sets that represents collective knowledge of gene function and interactions we refer to as pathways. The most common analysis utilizes biological processes from gene ontology[16] (GO) and molecular pathways from the Reactome database[17]. Depending on the hypothesis, these data may also include many other types of gene sets such as targets of transcription factors or microRNAs.

In the first step of ActivePathways, we derive an integrated gene list that for each input gene aggregates significance from multiple omics datasets. The integrated gene list is compiled by a fusion of gene significance from different omics datasets (i.e., evidence) using the Brown's extension[18] of the Fisher's combined probability test. The Brown's method considers dependencies between datasets and thus provides more conservative estimates of significance for genes that are supported by multiple similar omics datasets. The integrated input gene list is then ranked by decreasing significance and filtered using a lenient cut-off, designed to capture additional candidate genes with sub-significant signals while discarding the bulk of insignificant genes (unadjusted Brown $P_{gene} < 0.1$). In the second step, a pathway enrichment analysis is conducted on the integrated gene using a ranked hypergeometric test[10] and a collection of gene sets (i.e., biological processes, molecular pathways, and other gene annotations). The ranked hypergeometric test is designed to capture smaller pathways tightly associated with few top-ranking genes and also broader processes associated with larger subsets of input genes. The family-wise multiple testing correction method by Holm[19] is then applied across tested pathways to select the pathways significantly enriched in the integrated gene list ($Q_{pathway} < 0.05$). In the third step, we perform a similar analysis on the gene lists of individual omics datasets separately to determine the omics evidence supporting the integrative pathway analysis results determined in step 2. Importantly, the third step also highlights pathways that are only found through data integration and are not apparent in any single omics dataset separately, providing the added value of integrated analysis. Finally, the method provides input files for Enrichment Map[20] for visualizing resulting pathways with the corresponding omics evidence.

Integrating coding and non-coding drivers in 2658 cancer genomes. We performed an integrative pathway analysis of driver genes predicted in the PCAWG project based on somatic SNVs and indels. This analysis comprised 29 cancer patients cohorts of histological tumor types and 18 meta-cohorts combining multiple types of tumors, with 47 cohorts in total (Supplementary Table 1).
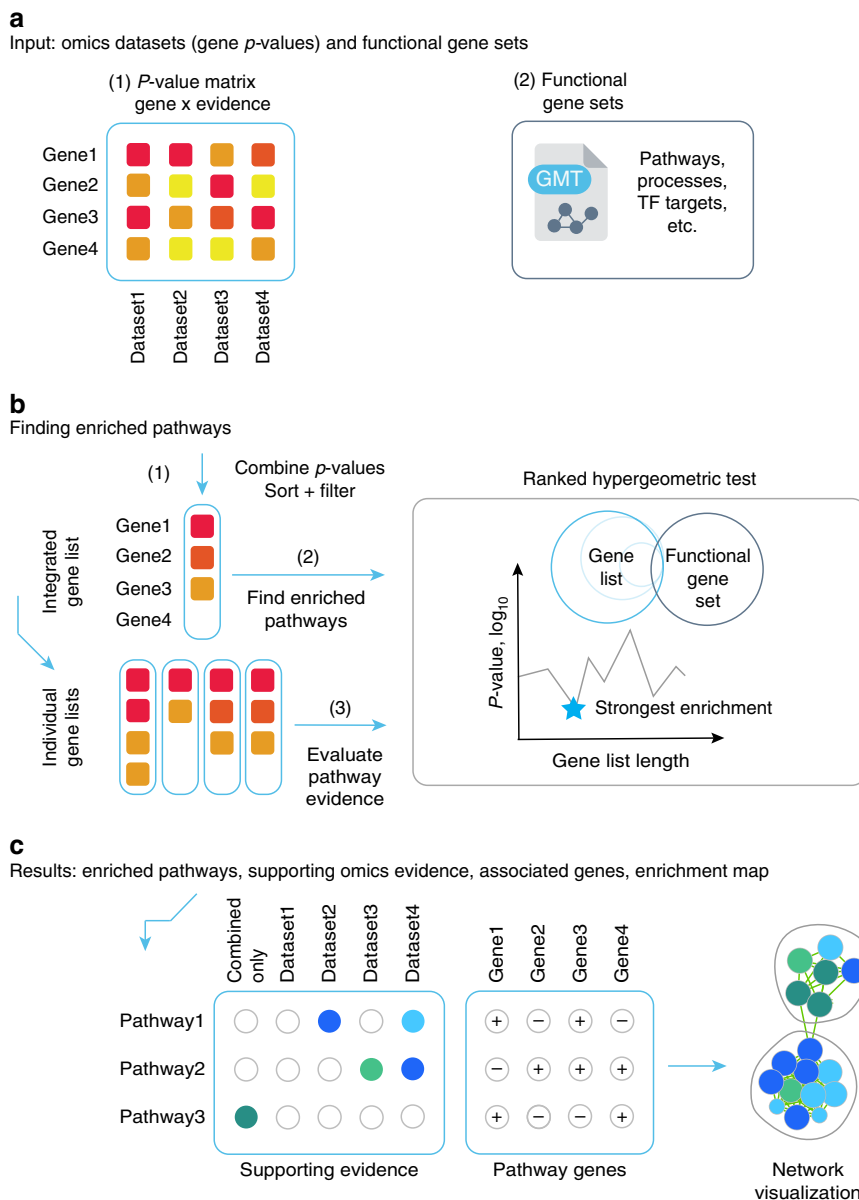
Fig. 1 Method overview. a ActivePathways requires as input (1) a matrix of gene *P*-values for different omics datasets, and (2) a collection of gene sets corresponding to biological pathways and processes. **b** In step (1), gene *P*-values are merged using the Brown procedure and filtered to produce an integrated gene list that combines evidence from datasets and is ranked by decreasing significance with a lenient threshold. In step (2), pathway enrichment analysis is conducted on the integrated gene list using the ranked hypergeometric test that determines the optimal level of enrichment in the ranked gene sub-list for every pathway. In step (3), separate gene lists are compiled from individual input datasets and analyzed for pathway enrichment using the ranked hypergeometric test, to find supporting evidence for each pathway from the integrative analysis. **c** ActivePathways provides a list of enriched pathways in the integrated gene list, the associated genes with significant Brown *P*-values, and annotations of evidence supporting each pathway. Results of ActivePathways are visualized as enrichment maps where nodes correspond to pathways and pathways with many shared genes are connected into networks representing broader biological themes.

ActivePathways identified at least one significantly enriched process or pathway in 89% of these cohorts (42/47, $Q_{pathway} <$ 0.05, ranked hypergeometric test) (Fig. 2a). We analyzed the evidence supporting predictions of enriched pathways: most cohorts showed enrichments in pathways supported by protein-coding mutations in genes (37/47 or 79%). This serves as a positive control since the majority of currently-known cancer driver genes have frequent protein-coding mutations.

Non-coding mutations in genes also contributed broadly to the discovery of frequently mutated biological processes and pathways: 24/47 cohorts (51%) showed significantly enriched pathways that were apparent when only analyzing non-coding driver

scores corresponding to UTRs, promoters or enhancers. The majority of PCAWG tumor cohorts (41/47 or 87%) revealed some frequently mutated pathways that were apparent when integrating coding and non-coding mutations however remained undetected when considering either coding or non-coding mutations separately, emphasizing the value of our integrative approach. As expected, cohorts with more patient tumor samples generated more significantly enriched pathways (Spearman $\rho =$ 0.74, $P = 2.3 \times 10^{-9}$; Supplementary Fig. 1), suggesting that larger datasets are better powered to distinguish rarely mutated genes involved in biological pathways and processes. Discovery of pathways enriched in non-coding mutations suggests that our
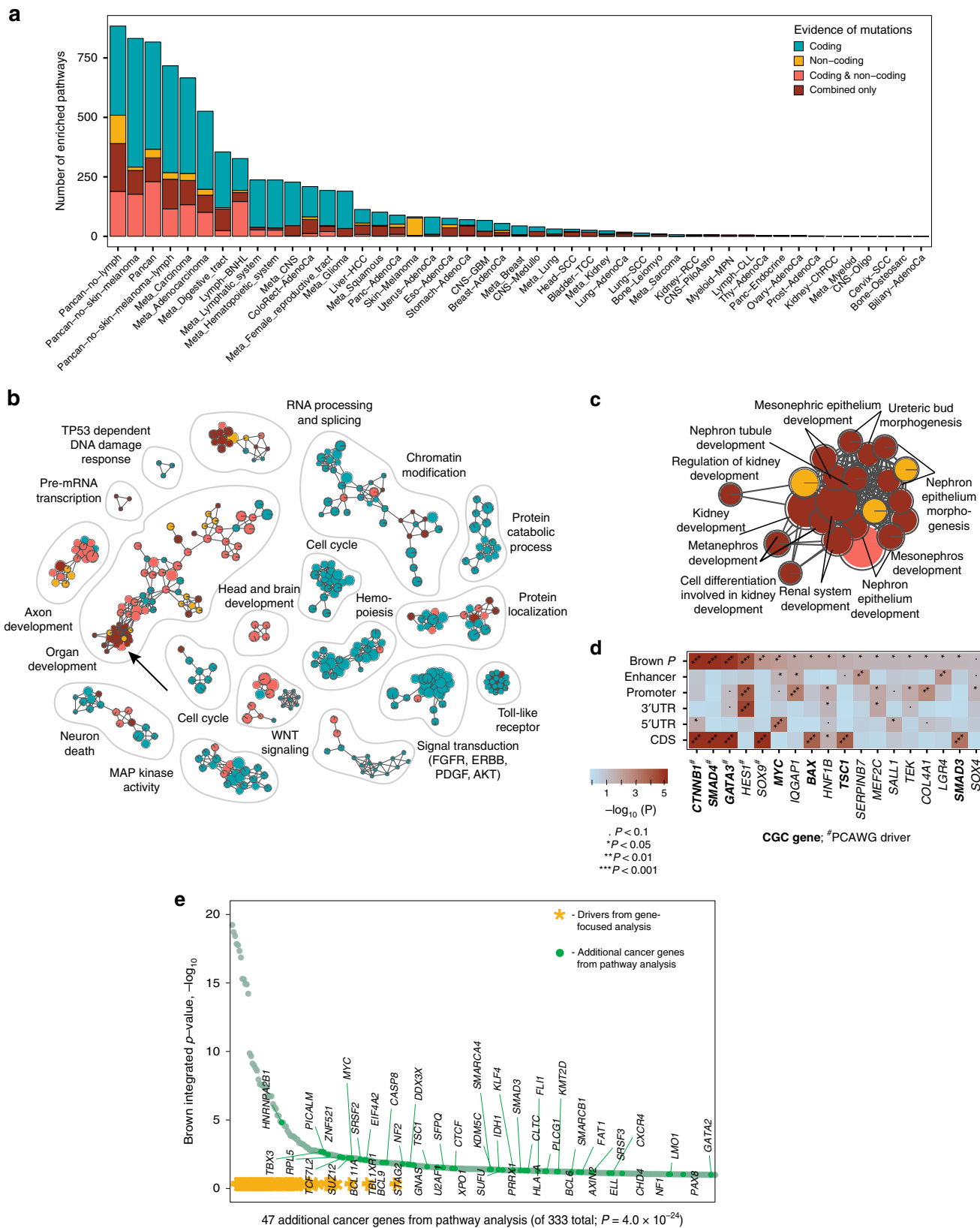
integrative pathway analysis is an attractive strategy for illuminating the dark matter of the non-coding cancer genome.

We studied the adenocarcinoma cohort of 1773 samples of 16 tumor types. Integrative pathway analysis highlighted 432 genes that were significantly enriched in 526 pathways ($Q_{pathway} < 0.05$) (Fig. 2b). As expected, the majority of pathways were only supported by genes with frequent coding mutations (328/526 or 62%). However, an additional set of 101 pathways (19%) was supported by both coding and non-coding gene mutations, 72 pathways (14%) were only apparent in the integrated analysis of both coding and non-coding mutations, and 25 (5%) were only enriched in non-coding mutations. Accumulation of these

Fig. 2 Pathway enrichment analysis of cancer driver genes with ActivePathways. a Bar plot shows number of significantly enriched pathways ($Q < 0.05$) among predicted driver genes with coding and non-coding mutations in the PCAWG dataset. The majority of pathways detected by ActivePathways are supported by protein-coding mutations, as expected (dark green bars), while non-coding mutations (orange, red) reveal additional pathways. Pathways shown in dark red are found only in the integrated gene list of coding and non-coding mutations but not in gene lists of individual mutation scores. **b** Enrichment map shows pathways enriched in frequently mutated genes in the adenocarcinoma cohort of 1773 tumors. Nodes in the network represent pathways and similar pathways with many common genes are connected. Groups of similar pathways are indicated. Nodes are colored by supporting evidence from coding and non-coding cancer mutations. Arrow indicates kidney developmental processes. **c** The group of enriched kidney developmental processes is apparent from integrated evidence of coding and non-coding mutations but is not found among coding or non-coding candidate genes separately. **d** Heatmap shows P-values of driver genes involved in kidney developmental processes, including driver genes found in the driver analysis (indicated with #) and additional genes only found in the pathway analysis. Top row shows merged P-values from the Brown procedure. Genes listed in the Cancer Gene Census (CGC) database are indicated in boldface letters. **e** Pathway analysis recovers most genes of the driver list from PCAWG (orange asterisks), as well as additional infrequently mutated genes apparent due to their pathway associations. Additional known cancer genes detected in the pathway analysis are listed (green dots) and occur more frequently than expected from chance alone.

---

individually infrequent non-coding mutations into relevant pathways and processes is apparent in our integrative analysis and remains undetected in a gene-focused analysis.

The major biological themes with frequent protein-coding mutations included hallmark cancer processes such as 'apoptotic signaling' (24 genes; $Q_{\text{pathway}} = 4.3 \times 10^{-5}$) and 'mitotic cell cycle' (8 genes; $Q_{\text{pathway}} = 0.0026$), and additional biological processes such as chromatin modification and RNA splicing that are increasingly recognized in cancer biology. Thus, ActivePathways captures the expected cancer pathways enriched in driver genes with protein-coding mutations as positive controls. In contrast to these solely protein-coding driver associations, a large group of developmental processes and signal transduction pathways was detected as enriched in both coding and non-coding mutations in genes; for example 'embryo development process' was supported by mutations in exons, 3'UTRs and gene promoters (68 genes; $Q_{\text{pathway}} = 2.9 \times 10^{-12}$), while the Reactome pathway 'repression of WNT target genes' was only apparent in the integrated analysis of coding and non-coding mutations but not in either dataset alone (5 genes, $Q_{\text{pathway}} = 0.016$). In summary, these data show that ActivePathways is a sensitive approach for integrating multi-omics signals such as coding and non-coding mutations, interpreting supporting omics evidence, and finding additional functional associations that are not apparent in any single input dataset.

Pathway-based prioritization of rarely mutated cancer genes. Pathway analysis can identify candidate genes that would otherwise remain undetected in gene-based analyses. ActivePathways enhances such discovery by integrating signals across multiple datasets. In the pathway analysis of coding and non-coding mutations in PCAWG, we focused on a group of processes involved in kidney development that were exclusively detected through the integration of coding and non-coding mutations (Fig. 2c, d). ActivePathways found 18 genes involved in these processes, only five of which were predicted as driver genes in the consensus driver analysis of the PCAWG project[13]. Additional known cancer genes included the oncogene *MYC* with 13 patients with 3'UTR mutations (driver P-value $P_{\text{UTR3}} = 4.8 \times 10^{-4}$), the transcription factor *SMAD3* of the TGF-β pathway with 14 patients with protein-coding mutations ($P_{\text{CDS}} = 4.0 \times 10^{-4}$) and the growth inhibitory tumor suppressor gene *TSC1* with 23 patients with protein-coding mutations ($P_{\text{CDS}} = 1.4 \times 10^{-4}$) as well as candidate cancer genes such as *IQGAP1* with ten patients with promoter mutations ($P_{\text{promoter}} = 8.2 \times 10^{-4}$) that encodes a signaling protein involved in cell motility and morphology. The additional genes remained below the FDR-adjusted significance cut-off in the gene-focused consensus driver analysis ($Q = 0.17$–0.62), however were found by ActivePathways due to their associations with kidney development. Thus ActivePathways can

exploit functional gene annotations and multiple omics signals to find further candidate genes that remain undetected in gene-focused analyses.

We evaluated the effects of our data integration strategy and examined all 333 pathway-associated candidate genes detected in the adenocarcinoma cohort (Fig. 2e). As expected, these included a considerable proportion of known cancer genes. First, as positive controls we found 60/64 significantly mutated genes that were also identified in the PCAWG consensus driver analysis[13]. In addition, we found a set of 47/333 known cancer genes annotated in the COSMIC Cancer Gene Census database[11] that were not detected in the driver analysis, significantly more than expected by chance alone (seven genes expected, Fisher's exact $P = 4.0 \times 10^{-24}$). Those included well-known cancer driver genes *MYC*, *IDH1*, *NF1*, and *BCL9*. ActivePathways was able to detect these additional genes for several reasons. First, the integrated gene list was filtered using a lenient statistical cut-off in ActivePathways ($P_{\text{gene}} < 0.1$) that allowed a long tail of 273/333 genes with less-frequent mutations to be detected through pathway associations. Second, certain genes were upgraded through the data fusion procedure as a single stronger P-value per gene was derived by combining multiple weaker P-values corresponding to the coding regions, promoters, UTRs, enhancers of the gene. This affected 17/333 pathway-associated genes including six known cancer genes (*HNRNPA2B1*, *STAG2*, *TCF7L2*, *SUZ12*, *CLTC*, and *ZNF521*), Thus, the integration procedure prioritized specific pathway-related genes compared to their original rankings in individual mutation datasets. However, the majority of genes showed reduced significance after the fusion and were excluded from the pathway analysis: 3543 genes had at least one significant P-value prior to data fusion ($P_{\text{gene}} < 0.1$) while 88% of these (3112) were considered insignificant following the Brown P-value combination step. In contrast, the majority of pathway-associated genes (220/333) showed improved rankings in the integrated gene list compared to their original rankings in individual input datasets. This formal combination of P-values across omics datasets is therefore more conservative than a naïve approach of selecting a top P-value for every gene. Thus, ActivePathways finds additional candidate genes that remain undetected in gene-by-gene analyses and are highlighted due to their multiple omics signals in pathways.

Integrating prognostic CNA and mRNA signals in breast cancer. To demonstrate an integrative analysis of patient clinical information with multiple types of omics data, we studied the pathways and processes associated with patient prognosis in breast cancer. We leveraged the METABRIC dataset[15] of 1780 breast cancer samples drawn from all four subtypes (HER2-enriched, basal-like, luminal-A, luminal-B) and evaluated all genes using three types of prognostic evidence. mRNA abundance

profiles were deconvolved between mRNA abundance levels in tumor cells (TC) and tumor-adjacent cells (TAC) using the ISOpure algorithm[21,22]. mRNA values were associated with patient survival using median dichotomization and log-rank tests. Gene copy number alterations (CNA) were included as the third type of evidence and associated with patient survival using log-rank tests.

ActivePathways identified 192 significantly enriched GO biological processes and Reactome pathways across the four subtypes of breast cancer, of which nine pathways were enriched in multiple cancer subtypes and 33 pathways were only apparent through the integrative pathway analysis but not in any of the CNA or mRNA datasets alone. The major findings enriched in prognostic signatures in breast cancer subtypes involved the processes and pathways of immune response, apoptosis, ribosome biogenesis and chromosome segregation (Fig. 3a).

Immune activity was associated with prognostic genes in basal-like and HER2-enriched breast cancers with significant enrichment of GO processes such as 'immune system development' ($Q_{basal} = 3.0 \times 10^{-4}$, 113 genes; $Q_{HER2} = 0.035$, 61 genes; ranked hypergeometric test) and 'lymphocyte differentiation' ($Q_{HER2} = 6.8 \times 10^{-4}$, 46 genes; $Q_{basal} = 8.4 \times 10^{-4}$, 45 genes). The majority of related genes were associated with improved patient prognosis upon increased mRNA abundance in tumor cells or tumor-adjacent cells, comprising 50/61 genes in the HER2-enriched subtype and 78/113 genes in the basal subtype (Fig. 3b). Interestingly, only a minority of these genes (10) were significant in both of the two breast cancer subtypes, suggesting different modes of immune activity in subtypes and emphasizing the power of our pathway-based approach. Basal-like breast cancers were associated with additional 67 terms involving immune response and blood cells, however no immune-related terms were enriched for luminal subtypes of breast cancers. Prognostic features of immune-related genes in HER2-enriched and basal-like breast cancers are well known[23,24]. Our pathway-based findings indicate that immune activity in breast tumor cells and in the surrounding microenvironment negatively affects tumor progression and improves prognosis.

Apoptosis was associated with patient prognosis in HER2-enriched and luminal-A breast cancers through enriched GO processes such as 'negative regulation of apoptotic process' ($Q_{HER2} = 0.030$, 122 genes; $Q_{luminalA} = 0.015$, 228 genes) and 'programmed cell death' ($Q_{HER2} = 0.015$, 125 genes; $Q_{luminalA} = 0.016$, 231 genes) (Fig. 3c). Interestingly, anti-apoptotic pathways were only detected in the integrative analysis and not in genomic and transcriptomic gene signatures separately. Among the genes negatively regulating apoptosis, DUSP1 provided the strongest prognostic signal in HER2-enriched breast cancers. This was apparent in the molecular stratification of samples by mRNA of tumor cells (log-rank $P_{TC} = 0.019$, HR = 1.5) and tumor-adjacent cells ($P_{TAC} = 8.3 \times 10^{-4}$, HR = 1.8) as well as gene copy number amplifications ($P_{CNA} = 9.8 \times 10^{-4}$, HR = 2.8) (Fig. 3d). DUSP1 encodes a phosphatase signaling protein of the MAPK pathway that is over-expressed in malignant breast cancer cells and inhibits apoptotic signaling[25]. HER2 over-expression is known to suppress apoptosis in breast cancer[26]. Anti-apoptotic signaling is a hallmark of cancer and expectedly associated with worse patient prognosis.

ActivePathways also identified prognostic pathway associations that were only apparent in single subtypes of breast cancer. For example, the prognostic genes for luminal-B subtype were enriched in processes related to 'chromosome segregation' ($Q_{luminalB} = 0.017$, 41 genes) that have been associated with worse outcome in breast cancer[27]. As another example, luminal-A breast cancers were associated with prognosis in ribosomal and RNA processing genes, such as 'ribosome biogenesis' ($Q_{luminalA} =$

6.9×10^{-10}, 60 genes) and 'rRNA metabolic process' ($Q_{luminalA} = 1.8 \times 10^{-13}$, 64 genes). Although not described specifically in the luminal-A subtype, ribosomal mRNA abundance has been shown to be prognostic in breast cancer as a marker of cell proliferation[28,29]. In summary, ActivePathways can be used for integrating clinical data with multiomics information of molecular alterations. Such analyses can provide leads for functional studies and biomarker development.
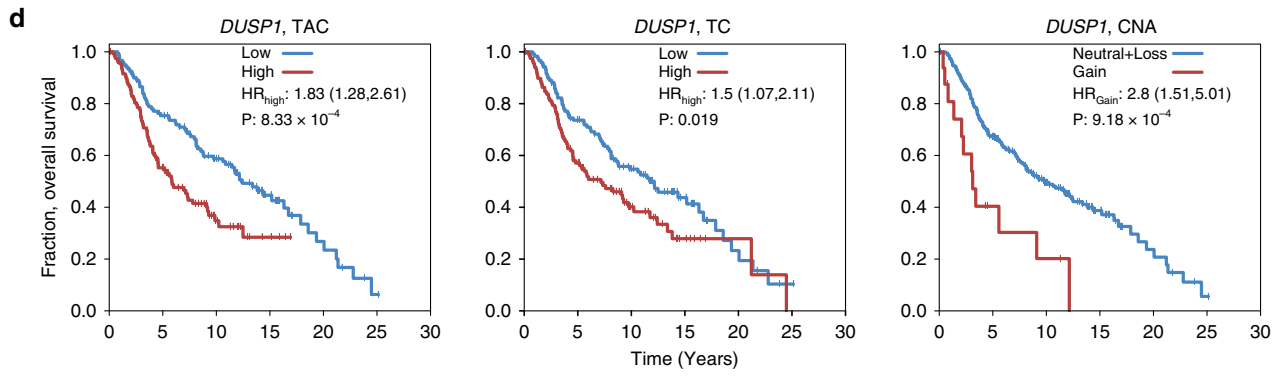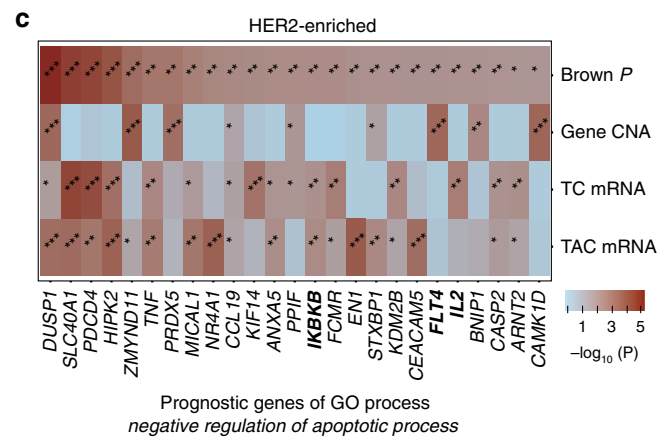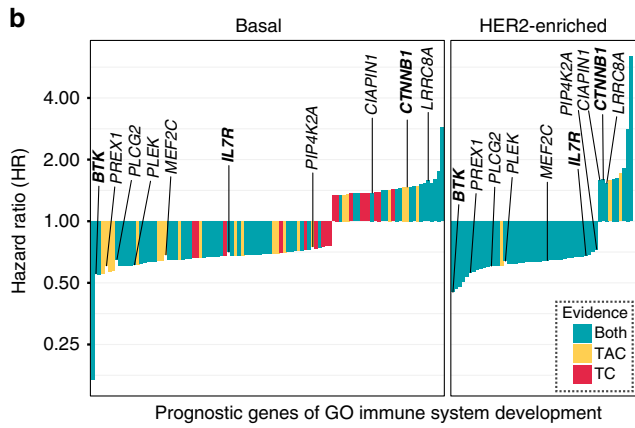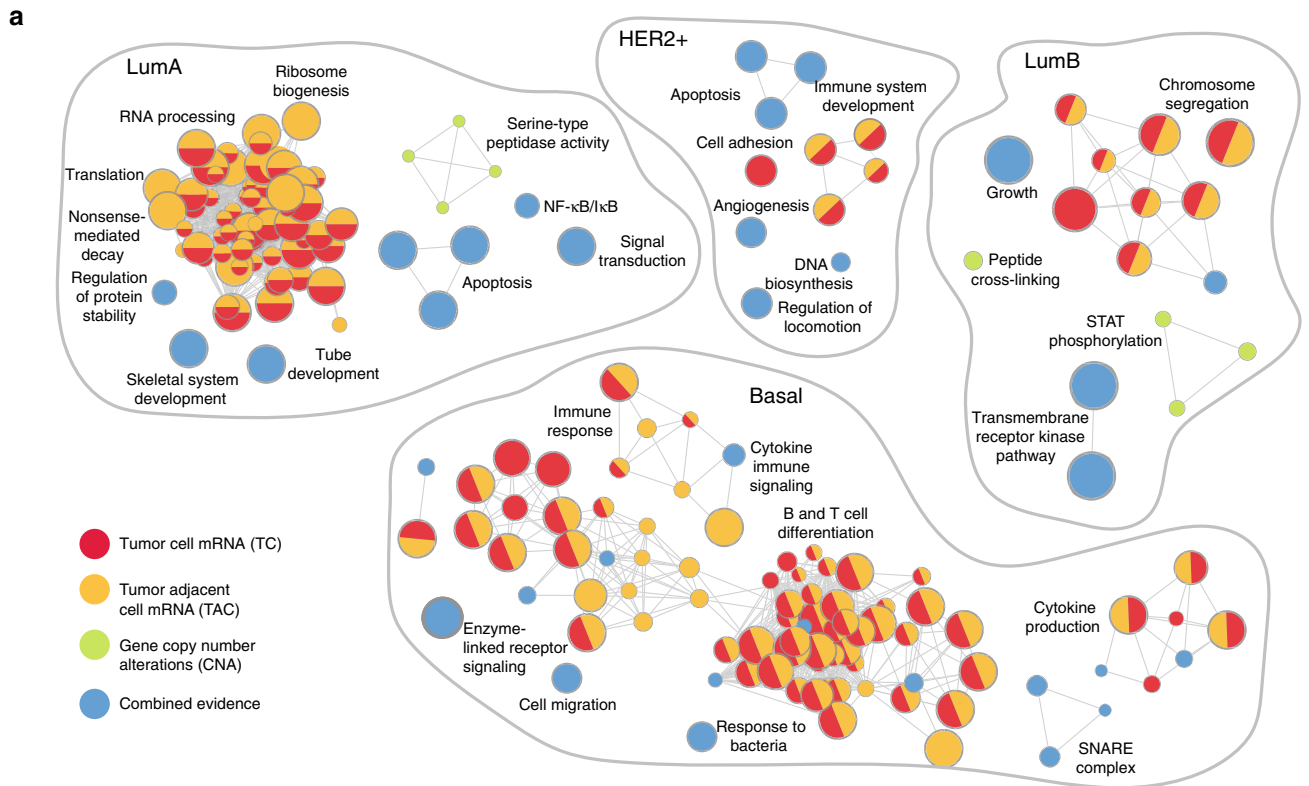
Interpreting co-expressed and DNA-bound targets of Hippo TFs. To demonstrate the use of ActivePathways for studying gene regulation, we analyzed transcriptomes of non-cancerous human tissues from the GTEx project[5]. We focused on the Hippo signaling pathway involved in organ size control, tissue homeostasis and cancer[30,31] and studied regulatory networks downstream of the two transcription factors (TFs) YAP and TAZ (encoded by YAP1 and WWTR1). YAP and TAZ are evolutionarily conserved master regulators of Hippo signaling in mammals that respond to intracellular and extracellular signals of cell–cell interactions, cell polarity, mechanical cues, G protein-coupled receptor signaling, and cellular energy status[30,31].

We performed an integrative pathway enrichment analysis of transcriptomics and epigenomics data of the two master regulators of the Hippo pathway. First, we predicted transcriptional target genes for YAP and TAZ (1898 and 1319 genes, respectively), using co-expression analysis and robust rank aggregation[32] over 9642 transcriptomes of 40 tissue types in GTEx ($Q < 0.05$). Second, we studied the set of 2356 target genes of YAP that have DNA-binding sites at gene promoters derived from a YAP ChIP-seq study[33] reanalyzed in the ReMap database[34] ($Q < 0.05$). The three gene lists with corresponding significance values were used as input to ActivePathways for the integrative analysis.

Integrative analysis of transcriptional and DNA-binding target genes of YAP/TAZ resulted in 225 significantly enriched GO processes and Reactome pathways ($Q_{pathway} < 0.05$) (Fig. 4a). The resulting pathways are expected in the context of Hippo signaling and included development and morphogenesis, cell motility, organization of actin cytoskeleton and cell–cell junctions, signal transduction pathways such as EGFR, Wnt, Robo, TGF-beta, rho GTPase, and others. ActivePathways highlighted 2066 genes in these pathways. We examined known members of the Hippo pathway and found 13 of 32 genes among pathway-associated genes (WWTR1, LATS1, WWC1, YWHAQ, YAP1, MAPK8, MAPK9, STK3, TNIK, TEAD3, MAP4K3, MINK1, TEAD1), more than expected from chance alone (four genes expected, $P = 3.6 \times 10^{-5}$, Fisher's exact test). We also compiled an extended list of 106 Hippo-related genes[30,31,35] and confirmed that these were enriched in pathway-associated genes (36/308 genes expected, $P = 9.1 \times 10^{-26}$, Fisher's exact test) (Fig. 4b). A large fraction of pathways (55 or 24%) was only identified in the integrated analysis and not in any input dataset alone, underlining the advantage of detecting significantly enriched pathways across multiple complementary omics datasets. Similarly, Hippo-related genes were either supported by ChIP-seq data alone (54 genes), mRNA data alone (30 genes), or both by mRNA and ChIP-seq data (22 genes), concordant with the notion that RNA-seq and ChIP-seq show limited agreement regarding TF target genes and provide complementary insights into gene regulation. Thus, our integrative analysis of co-expression networks and TF–DNA interactions of Hippo master regulators expectedly converges on Hippo-related genes and pathways.

In addition to GO terms and pathways, ActivePathways can be used to interpret omics data with other classes of functional gene sets such as TF target genes. To further elucidate gene regulatory

a

LumA — Ribosome biogenesis, RNA processing, Translation, Nonsense-mediated decay, Regulation of protein stability, Skeletal system development, Tube development, Serine-type peptidase activity, NF-κB/IκB, Signal transduction, Apoptosis

HER2+ — Apoptosis, Immune system development, Cell adhesion, Angiogenesis, DNA biosynthesis, Regulation of locomotion

LumB — Chromosome segregation, Growth, Peptide cross-linking, STAT phosphorylation, Transmembrane receptor kinase pathway

Basal — Immune response, Cytokine immune signaling, B and T cell differentiation, Enzyme-linked receptor signaling, Cell migration, Response to bacteria, Cytokine production, SNARE complex

Legend:
- Tumor cell mRNA (TC)
- Tumor adjacent cell mRNA (TAC)
- Gene copy number alterations (CNA)
- Combined evidence

b

Basal; HER2-enriched

Prognostic genes of GO immune system development

Hazard ratio (HR): 4.00, 1.00, 0.50, 0.25

Genes labeled: BTK, PREX1, PLCG2, PLEK, MEF2C, IL7R, PIP4K2A, CIAPIN1, CTNNB1, LRRC8A

Evidence: Both, TAC, TC

c

HER2-enriched

Prognostic genes of GO process *negative regulation of apoptotic process*

Rows: Brown *P*, Gene CNA, TC mRNA, TAC mRNA

Genes: DUSP1, SLC40A1, PDCD4, HIPK2, ZMYND11, TNF, PRDX5, MICAL1, NR4A1, CCL19, KIF14, ANXA5, PPIF, IKBKB, FCMR, EN1, STXBP1, KDM2B, CEACAM5, FLT4, IL2, BNIP1, CASP2, ARNT2, CAMK1D

$-\log_{10}(P)$: 1, 3, 5

d

DUSP1, TAC — Low, High; HR_high: 1.83 (1.28,2.61); P: 8.33×10^{-4}

DUSP1, TC — Low, High; HR_high: 1.5 (1.07,2.11); P: 0.019

DUSP1, CNA — Neutral+Loss, Gain; HR_Gain: 2.8 (1.51,5.01); P: 9.18×10^{-4}

Fraction, overall survival vs Time (Years)

networks downstream of YAP/TAZ, we considered potential enrichments of DNA-binding target genes of 161 TFs profiled in ChIP-seq studies of the ENCODE project[36]. We found a regulatory network of 17 TFs and 1426 target genes enriched in the YAP/TAZ regulome ($Q < 0.05$, ranked hypergeometric test) (Fig. 4c). These included the DNA-binding targets of the master

regulators of pluripotent stem cells[37] NANOG (208/774 target genes, $Q = 1.3 \times 10^{-6}$) and POU5F1 (107/406 target genes, $Q = 7.3 \times 10^{-11}$). This finding is in agreement with the role of Hippo pathway activity in stem cell regulation[38]. The regulatory network was significantly enriched in 50 Hippo-related genes (25 expected, $P = 1.2 \times 10^{-7}$, Fisher's exact test) and six core Hippo genes (two

Fig. 3 Prognosis-associated pathways in four molecular subtypes of breast cancer. a Enrichment maps of prognostic pathways and processes were found in an integrative analysis of mRNA abundance in tumor cells (TC), tumor-adjacent cells (TAC), and gene copy number alterations (CNA) of the METABRIC dataset. Multicolored nodes indicate pathways that were prognostic according to several types of molecular evidence. Blue nodes indicate pathways that were only apparent through merging of molecular signals. **b** Hazard ratios (HR) of prognostic genes related to immune system development in basal and HER2-enriched subtypes of breast cancer. Strongest HR value of TC, TAC is shown. Genes commonly found in basal and HER2-enriched tumors are shown. Known cancer genes are shown in boldface. **c** Heatmap shows genes, corresponding log-rank P-values, and merged Brown P-values related to the GO process 'negative regulation of apoptotic process' that was found by integrating prognostic omics data in HER2-enriched breast cancer. **d** Kaplan–Meier plots show the strongest prognostic signal related to apoptotic signaling, the phosphatase *DUSP1* that significantly associates with reduced patient survival through increased tumor-adjacent mRNA level (left), increased tumor mRNA level (center) and gene copy number amplification (right). Log-rank P-values are shown.

expected, $P = 0.030$; *WWTR1, VGLL4, TNIK, MAPK8, MOB1A, LATS1*), similarly to the pathway-based analysis above. However, the two analyses revealed distinct genes: 886 genes were commonly found by both analyses, 1180 genes were found only in the pathway-based analysis, and 540 genes were found only in the TF-based analysis (Fig. 4d). Thus, integrative enrichment analysis of TF target genes provides complementary information to GO terms and pathways. In summary, this analysis highlights genes and pathways related to Hippo signaling in human tissues and demonstrates the use of ActivePathways for studying gene regulatory networks across complementary omics datasets and technology platforms.

Evaluating the robustness and sensitivity of ActivePathways. We carefully benchmarked ActivePathways using the dataset of cancer driver genes predicted by PCAWG[13]. First, we compared the performance of ActivePathways with six methods used in the PCAWG pathway and network analysis working group[14] (Hierarchical HotNet[39,40], SSA−ME[41], NBDI[42], induced subnetwork analysis[40], CanIsoNet[43], hypergeometric test). These diverse methods used molecular interaction networks, functional gene sets and/or transcriptomics data to analyze the PCAWG pan-cancer dataset of predicted cancer driver genes. A subsequent consensus analysis defined pathway-implicated driver gene lists with protein-coding and non-coding mutations, based on a majority vote of the pathway and network analysis methods[14]. ActivePathways recovered these consensus gene lists with the highest accuracy: 100% of coding driver genes (87/87) and 85% of non-coding candidate genes (79/93) were detected (Fig. 5a). Thus ActivePathways agreed the most with the ensemble of several other methods in prioritizing known and candidate cancer driver genes using pathway and network context.

We also compared the performance of ActivePathways with a standard approach of pathway enrichment analysis that considers a single statistically-filtered gene list using a ranked hypergeometric test[1,10] (Fig. 5b). To this end, we analyzed individual gene lists of protein-coding and non-coding drivers of the PCAWG adenocarcinoma cohort using multiple gene selection thresholds ($Q < 0.05$, $Q < 0.1$, $Q < 0.25$, $P < 0.1$). ActivePathways showed increased sensitivity of pathway enrichment analysis compared to the standard approach, in particular for pathways involving non-coding mutations that were not prominently represented in any single gene lists. Even compared to the analysis employing the most lenient gene selection filter ($P < 0.1$), ActivePathways identified 72 additional pathways that were only apparent through the integration of coding and non-coding mutations and remained undetected in the analysis of individual datasets. Thus, our method provides additional information to common approaches that focus on single gene lists.

We evaluated the robustness of ActivePathways to parameter variations and missing data. We varied the parameter P_{gene} that determines the ranked gene lists used in the pathway enrichment analysis (default threshold $P_{gene} < 0.1$). The majority of PCAWG cohorts (40/47 or 85%) retrieved significantly enriched pathways

even with a considerably more stringent threshold ($P_{gene} < 0.001$), however 67% fewer pathways were found compared to the default threshold in the median cohort (Supplementary Fig. 2). We then evaluated the robustness of ActivePathways to missing data by randomly removing subsets of driver scores from the initial dataset. Even when removing 50% of gene driver P-values with $P < 0.001$, the majority of cohorts (37/47 or 79%) had at least one significantly enriched pathway, however 66% fewer pathways were found on average (Supplementary Fig. 3).

We evaluated the expected false positive rates of ActivePathways. We tested 1,000 simulated datasets for each of 47 patient cohorts and expectedly found no significant pathways in 92% of these simulations (Supplementary Fig. 4). Simulated data were obtained by randomly reassigning P-values of driver predictions to different genomic elements, a conservative approach that disrupts pathway annotations of genes while retaining the presence of strong P-values observed in the real data. The median family-wise false discovery rate of ActivePathways computed across cohorts (7.2%) slightly exceeded the applied multiple testing correction ($Q < 0.05$). Higher rates were observed in cohorts including melanoma tumors, potentially due to abundant promoter mutations caused by impaired nucleotide excision repair in protein-bound genomic regions[44]. We evaluated quantile-quantile (QQ) plots of pathway-based ranked hypergeometric P-values from ActivePathways and found that these often deviated from the expected uniform distribution and appeared statistically inflated (Supplementary Fig. 5). However, P-values derived from simulated gene scores showed no inflation in our simulations. Anticipating that the most significant cancer driver genes associated with protein-coding mutations, we performed partial simulations. We constructed datasets with simulated gene P-values for protein-coding mutations and true P-values for non-coding mutations. As expected, partially simulated datasets showed a lesser extent of P-value inflation, suggesting that highly significant known cancer genes are responsible for inflation due to their involvement in many pathways. Statistical testing of highly redundant pathways and processes violates the independence assumption of statistical tests and multiple testing procedures, a known caveat of pathway enrichment analysis[1], which likely explains the observed distribution of significance values of our method. Collectively, these benchmarks show that ActivePathways is a sensitive and robust method for detecting significantly enriched pathways and processes through integrative analysis of multi-omics data.

Discussion
Integrative pathway enrichment analysis helps distill thousands of high-throughput measurements to a smaller number of pathways and biological themes that are most characteristic of the experimental data at hand, ideally leading to mechanistic insights and candidate genes for follow-up studies. In particular, a joint analysis of complementary datasets often leads to insights that are unavailable in any particular dataset. ActivePathways provides a generally available framework for systematically prioritizing genes

and pathways across multiple omics datasets that utilizes fusion of gene significance. This allows us to identify pathways and processes that stand out when the data are combined, yet are not apparent in any single analyzed dataset. The use of

ActivePathways is demonstrated in the three case studies above. In our example of cancer driver discovery, pathway analysis provides complementary insights to gene-focused driver discovery as it also focuses on sub-significant genes that are clustered

Fig. 4 Integrative pathway enrichment analysis of Hippo target genes across human tissues. a Enrichment map of GO processes and Reactome pathways enriched in the target genes of transcriptional regulators YAP and TAZ of the Hippo pathway. Co-expressed target genes of YAP and TAZ across normal human tissues of the GTEx dataset (pathways are shown in green and yellow, respectively) and DNA-binding target genes of YAP from ChIP-seq experiments (shown in blue) were analyzed. Pathways only found through the integration of ChIP-seq and RNA-seq data are shown in red. **b** Euler diagram shows 106 Hippo-related genes that were significantly enriched in the detected pathways and supported by a combination of signals in RNA-seq and ChIP-seq datasets. Core Hippo genes detected in the analysis are listed. **c** Regulatory network of 17 TFs and 1,426 target genes detected in the ActivePathways analysis of gene sets representing transcription factor target genes. Transcription factors with enriched target genes in the YAP/TAZ regulome are shown in multi-colored circles. Target genes are colored by increasing statistical significance (turquoise to red). **d** Integrative analysis of pathways and GO processes is complementary to the analysis of transcription factor targets. Euler diagram shows total number of pathway-associated identified in the analysis of GO and Reactome terms (left) and TF target genes from ENCODE (right). Numbers of Hippo-related genes are shown in brackets.
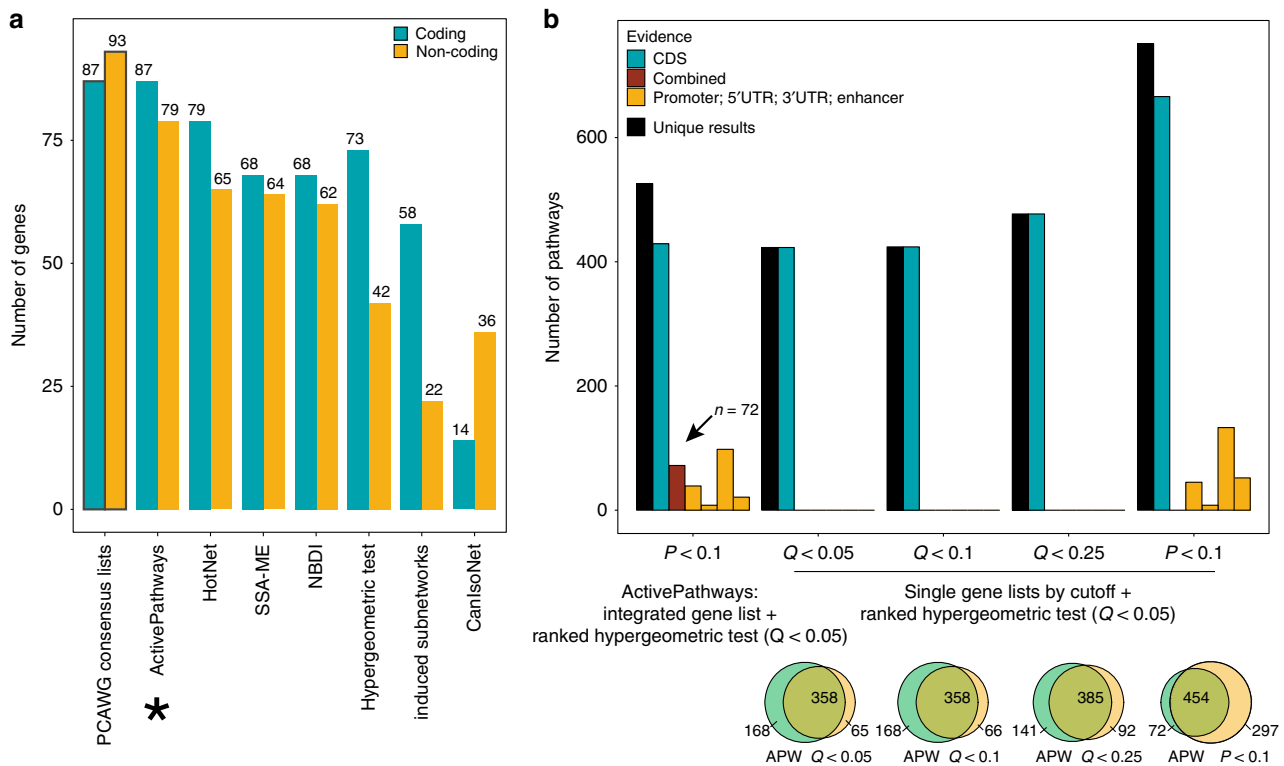

Fig. 5 Benchmarking of ActivePathways. a Comparison of ActivePathways with six additional pathway and network analysis methods used in the PCAWG pathway and network consensus analysis. ActivePathways best recovers the consensus lists of pathway-implicated driver genes with coding and non-coding mutations (indicated by asterisk). The consensus lists are shown in the leftmost bars of the plot and have been compiled through a majority vote of the seven methods in the PCAWG pathway and network analysis working group. **b** Comparison of ActivePathways (leftmost bars) and common pathway enrichment analysis using multiple significance cut-offs of PCAWG gene lists with protein-coding and non-coding mutations. ActivePathways shows increased sensitivity of pathway analysis even at the most lenient gene list significance cut-offs and recovers additional pathways only detected through integration of multiple datasets (dark red).

into relevant biological processes and carry coding and non-coding mutations. In the integrative analysis of molecular alterations in breast cancer subtypes, we find a spectrum of genes and pathways whose molecular signatures in the tumors or the microenvironment have potential prognostic significance. A subset of these findings, such as anti-apoptotic signaling, is only apparent through data integration. In the final example, we use ActivePathways to associate the co-expression and DNA-binding networks of Hippo master regulators to downstream pathways and processes with multi-omics evidence.

Our general pathway analysis strategy is applicable to diverse kinds of datasets where well-calibrated P-values are available for the entire set of genes or proteins. A multi-omics study may quantify genes using a series of genomic, transcriptomic and proteomic experiments and compute corresponding P-values. Data from epigenomic experiments and genome-wide association

studies can be also analyzed after signals have been appropriately mapped to genes, for example by identifying ChIP-seq peaks in gene promoters similarly to our GTEx analysis. Clinical and phenotypic information of patients can be also included through association and survival statistics, as shown in our analysis of prognostic signatures of breast cancer subtypes. Our method is expected to work with raw, unadjusted P-values and also with Q-values adjusted for multiple testing, however it is primarily intended for unadjusted P-values for increased sensitivity. ActivePathways conducts multiple testing correction at the pathway level and reports significantly enriched pathways at a family-wise error rate cutoff, regardless of the gene-specific multiple testing correction applied upstream. Quantification of genes and proteins through P-values is more robust than quantification through their abundance measures such as counts or fold-changes. P-values provided to ActivePathways need to be computed using dedicated

methods for individual omics platforms such that inherent biases in the data are accounted for prior to pathway analysis. In our example of cancer driver discovery, appropriately computed P-values of driver gene predictions from the PCAWG project[13] account for confounding factors of somatic mutations, such as gene sequence length and nucleotide content, mutation signatures active in different types of tumors[45] and biological correlates of mutation frequency such as transcription and replication timing[46]. On the one hand, considering all such variations directly in the pathway enrichment analysis would require substantially more complex models. On the other hand, directly analyzing pathways using simpler metrics (such as mutation counts or frequencies) would propagate any upstream biases to the pathway enrichment analysis and cause challenges with false positives and data interpretation. Thius, given appropriately derived P-values for genes, proteins and other molecules, ActivePathways can be applied to a wide range of analyses.

Our method comes with important caveats. First, we only evaluate genes and proteins annotated in pathway databases. Such databases have variable coverage, rely on frequent data updates[47] and may miss sparsely annotated candidate genes. The most general type of pathway enrichment analysis considers biological processes and molecular pathways however many kinds of gene sets available in resources such as MSigDB[48] can be used to expand the scope of ActivePathways. Second, pathway information is highly redundant and analyses of rich molecular datasets often result in many significant results reflecting the same underlying pathway. We address this redundancy by visualizing and summarizing pathway results as enrichment maps[1,20] that summarize multiple similar pathways and processes into general biological themes. Statistical inflation of results accompanied by biological redundancy is addressed by a stringent multiple testing correction. Third, the analysis treats pathways as gene sets and does not consider interactions of genes in pathways. This simplified strategy allows us to consider a wider repertoire of pathways and processes as reliable mechanistic interactions are often context-specific and limited to a small subset of well-studied signaling pathways. Several advanced methods such as HotNet[39], PARADIGM[49], and GeneMania[50] model pathways through gene and protein interactions.

Translation of discoveries into improved human health through actionable mechanistic insights, biomarkers, and molecular therapies is a long-standing goal of biomedical research. For example, next-generation cancer genomics projects such as ICGC-ARGO aim to collect multi-omics datasets with detailed clinical profiles that will present new challenges for pathway and network analysis techniques. In summary, ActivePathways is integrative pathway analysis method that improves systems-level understanding of cellular organization in health and disease.

Methods

Integrated and evidence-based gene lists. The main input of ActivePathways is a matrix of P-values where rows include all genes of a genome and columns correspond to evidence from omics datasets. To interpret multiple omics datasets, a combined P-value is computed for each gene using a data fusion approach, resulting in an integrated gene list. The integrated gene list is computed by merging all P-values of a given gene into one combined P-value using the Brown's extension[18] of the Fisher's combined probability test that accounts for overall covariation of P-values from different sources of evidence. The integrated gene list of Brown P-values is then ranked in order of decreasing significance and filtered using a lenient threshold (unadjusted $P < 0.1$ by default). Evidence-based gene lists representing different omics datasets are based on ranked P-values from individual columns of the input matrix and filtered using the same significance threshold.

Statistical enrichment of pathways. Statistical enrichment of pathways in ranked lists of candidate genes is carried out with the ranked hypergeometric test. The test considers one pathway gene set at a time and analyzes increasing subsets of input genes from the top of the ranked gene list. The same procedure is used for

integrated and evidence-based gene lists. At each iteration, the test computes the hypergeometric enrichment statistic and P-value for the set of genes shared by the pathway and top sub-list of the input gene list. For optimal processing speed, only gene lists ending with a pathway-related gene are considered. The ranked hypergeometric statistic selected the input gene sub-list that achieves the strongest enrichment and the smallest P-value as the final result for the given pathway, as:

$$\left(P_{\text{pathway}}, G\right) = \{\min, \text{argmin}_n\} \sum_{x=k}^{\min(n,K)} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},$$

where P_{pathway} stands for the hypergeometric P-value of the pathway enrichment at the optimal sub-list of the significance-ranked candidate genes, G represents the length of the optimal sub-list, i.e., the number of top genes from the input gene list, N is the number of protein-coding genes with annotations in the pathway database, i.e., in Gene Ontology and Reactome, K is the total number of genes in a given pathway, n is the number of genes in a given gene sub-list considered, and k is the number of pathway genes in the considered sub-list. For a conservative estimate of pathway enrichment, we consider as background N the universe of genes contained in input gene sets (terms from pathway databases and ontologies) rather than the complete repertoire of protein-coding genes. To obtain candidate genes involved in the pathway of interest, we intersect pathway genes with the optimal sub-list of candidate genes. The ranked hypergeometric P-value is computed for all pathways and resulting P-values are corrected for multiple testing using the Holm-Bonferroni method of family-wise error rate (FWER)[19]. Significant pathways are reported by default ($Q < 0.05$).

Evaluating omics evidence of enriched pathways. Each evidence-based gene list derived from a single omics dataset is also analyzed for enriched pathways with the ranked hypergeometric test. Pathways found in the integrated gene list are then labeled for supporting evidence in the case they are also found as significant in any evidence-based gene list. A pathway is considered to be found only through data integration and labeled as combined-only if it is identified as enriched in the integrated gene list but not identified as enriched in any of the evidence-based gene lists at equivalent significance cut-offs (default $Q < 0.05$). Each detected pathway is additionally annotated with pathway genes apparent in the optimal sub-list of candidate genes separately for the integrated gene list and each evidence-based gene list.

Pathways and processes. We used gene sets corresponding to biological processes of Gene Ontology[16] and molecular pathways of the Reactome database[17] downloaded from the g:Profiler web server[10] on Jan 26th 2018. Large general gene sets with more than a thousand genes and small specific gene sets with less than five genes were excluded to avoid statistical inflation of large gene lists and interpretation challenges with very small lists.

Enrichment map visualization. ActivePathways creates input files for the EnrichmentMap app[20] of Cytoscape[51] for network visualization of similar pathways and their coloring according to supporting evidence. Enrichment maps for adenocarcinoma driver mutations, prognostic quantification of molecular alterations of breast cancer, and transcriptional networks downstream of Hippo signaling were visualized with stringent pathway similarity scores (Jaccard and overlap combined coefficient 0.6) and manually curated for the most representative groups of similar pathways and processes. Singleton pathways that were redundant with larger groups of pathways were merged with the latter or discarded. Coloring of pathways in the adenocarcinoma enrichment map was rearranged by merging colors of pathways supported by non-coding mutation scores of promoters, enhancers and/or UTRs into one group.

Coding and non-coding mutations of the PCAWG dataset. We used ActivePathways to analyze driver predictions of coding and non-coding mutations across white-listed 2583 whole cancer genomes of the ICGC-TCGA PCAWG project. P-values of driver predictions were computed separately for protein-coding sequences, promoters, enhancers and untranslated regions (UTR3, UTR5 across multiple subsets of samples representing histological tumor types and pan-cancer cohorts as reported in the PCAWG driver discovery study[13]. We used gene-enhancer mapping predictions provided by PCAWG, excluded enhancers with more than five target genes, and selected the most significant enhancer for each gene, if any. Unadjusted P-values for coding sequences, promoters, enhancers and UTRs were compiled as input matrices and analyzed as described above. Missing P-values were conservatively interpreted as ones. Results from ActivePathways were validated with two lists of cancer genes. Predicted drivers from the PCAWG consensus analysis[13] were selected as statistically significant findings ($Q < 0.05$) following a stringent multiple testing correction spanning all types of elements (exons, UTRs, promoter, enhancers). The curated list of known cancer genes was retrieved from the COSMIC Cancer Gene Census (CGC) database[11]. One-tailed Fisher's exact tests were used to estimate enrichment of these genes in our results, using all human protein-coding genes as the statistical background set.

Prognostic CNA and mRNA signals in breast cancer. ActivePathways was used to evaluate prognostic pathways in breast cancer subtypes. Multiple types of omics data were used for an integrative analysis: mRNA gene expression data and gene copy number alteration (CNA) data of the were derived from the METABRIC cohort of 1991 patients with a single primary fresh frozen breast cancer specimen each[15]. Curtis et al.[15] classified the patients into the intrinsic breast cancer subtypes using the PAM50 mRNA-based classifier[52] resulting in 330 basal-like breast cancers, 238 HER2-enriched breast cancers, 721 luminal-A breast cancers, 491 luminal-B breast cancers. Using these data, we computationally deconvolved tumor cell (TC) mRNA and tumor adjacent cell (TAC) mRNA abundance levels from the bulk profiled specimens. TC mRNA was deconvolved using the ISOpure[21] method release 2010b in MATLAB. TAC mRNA abundance profiles were computed using the ISOpure.calculate.tac function from the R package ISOpureR[22] v1.1.2. The deconvolution analyses were performed independently for each breast cancer subtype. The mRNA univariate survival analysis was conducted as follows. For each gene, patients were dichotomized based on mRNA abundance. Dichotomization was either based on the median mRNA abundance for that gene or a fixed value of 6.5. Based on the mRNA abundance distribution of genes on the Y chromosome in female samples, the value 6.5 was estimated as the threshold for noise for non-expressed genes. Median dichotomization was used if the median was above 6.5 or if there were no events in one of the groups when dichotomizing based on 6.5. The high and low groups based on mRNA abundance were compared by univariate log-rank tests for overall patient survival. TC and TAC mRNA abundance values were evaluated independently. Survival modeling was performed in the R statistical environment (v3.4.3) using the survival package (v2.42–3). The CNA univariate survival analysis was conducted as follows. For each gene, we assessed whether more gains or losses were apparent. The copy number status with a higher count was subsequently used to separate patients into two groups: those with the chosen copy number status and the remaining patients. The two groups were then used for overall survival modeling with log-rank tests.

Hippo pathway target genes in mRNA and ChIP-seq data. This analysis included two types of omics data, mRNA abundance measurements from RNA-seq experiments and transcription factor DNA-binding measurements from chromatin immunoprecipitation sequencing (ChIP-seq) experiments. The RNA-seq dataset of human tissues was downloaded from GTEx v7 data portal (https://www.gtexportal.org/home/). The dataset included transcript abundance values of 21,518 protein-coding genes in 11,688 samples across 53 tissues. Tissues with less than 25 available samples and low gene expression (mean TPM < 1.0; transcripts per million) were excluded from further analysis, resulting in 40 tissues and 9672 samples with mRNA abundance profiles of 19,025 genes. Transcriptional target gene lists for the master transcription factors YAP and TAZ (encoded by *YAP1, WWTR1*) were predicted separately in the following two steps. First, we computed pairwise Pearson correlations between a given TF and all other genes within a tissue of interest and ranked these by the significance *P*-value of positive correlations. Second, the resulting ranked gene lists were then aggregated into one master target gene list across the GTEx tissues using the robust rank aggregation (RRA) method[32] with default parameters ($Q_{gene} < 0.05$). FDR-adjusted values of genes from RRA for YAP and TAZ were used as the first and second evidence for input of ActivePathways, respectively. For DNA-binding targets of YAP, we retrieved ChIP-seq binding sites in three cell lines (CCLP1, MSTO, and HUCCT1) from an earlier study[33] that were reprocessed in the ReMap database[34]. Binding sites were filtered by statistical significance (Q < 0.05) and mapped to gene promoters of the human genome (hg19) using gene promoters defined in the PCAWG driver analysis[13]. If a promoter had multiple binding sites, the site with the strongest FDR-value was selected as a representative site for that gene. FDR-adjusted values of genes with YAP DNA-binding sites were used as the third evidence for input of Active-Pathways. Significantly enriched pathways among the putative target genes of YAP and TAZ were subsequently detected using ActivePathways. We compiled a list of 308 Hippo-related genes from the KEGG pathway database[35] and two recent review papers[30,31]. To validate the analysis, we tested the overall sets of genes identified by ActivePathways for enrichment of Hippo-related genes using Fisher's exact tests. The Hippo analysis was conducted separately for two collections of functional gene sets. First, we tested GO biological processes and Reactome pathways similarly to analyses described above. Second, we used gene sets corresponding to transcription factor binding sites (TFBS) derived from the ENCODE project[36].

Method benchmarking. We benchmarked ActivePathways using multiple approaches, including simulated datasets of *P*-values, method parameter variations, and partial replacement of *P*-values scores with insignificant *P*-values. Benchmarking was performed on the PCAWG predictions of cancer driver genes with coding and non-coding mutations. To evaluate the false discovery rate of Active-Pathways, we created simulated datasets by reassigning all observed *P*-values to random genes and their genomic elements. Simulations were conducted separately for different tumor cohorts. One thousand simulated datasets were analyzed with ActivePathways and those with at least one significantly detected pathway counted towards false discovery rates. A separate set of simulations maintained the positions of non-coding *P*-values among genes and randomly reassigned *P*-values

corresponding to protein-coding mutations, expectedly leading to a reduction in detected pathways as the PCAWG datasets primarily included strong *P*-values for genes with frequent protein-coding mutations. Quantile-quantile analysis and QQ-plots were used to compare *P*-value distributions of pathways discovered from true *P*-values, partially shuffled P-values (true non-coding and shuffled protein-coding *P*-values), and fully shuffled *P*-values. To evaluate robustness of ActivePathways, we randomly replaced a fraction of significant driver *P*-values in the true dataset (*P* < 0.001) with insignificant *P*-values (*P* = 1). We tested different fractions of missing values (10, 25, and 50%) across a thousand datasets of randomly selected missing data points. We concluded that most PCAWG cohorts included significantly enriched pathways even with large fractions of missing data. To further evaluate robustness, we tested different values of the Brown *P*-value threshold used to select the integrated gene list for pathway enrichment analysis. The default parameter value ($P_{gene} < 0.1$) was compared to alternative values (0.001, 0.01, 0.05, and 0.2). We concluded that ActivePathways found enriched pathways in most tumor cohorts even at more stringent gene selection levels.

Data availability

Somatic and germline variant calls, mutational signatures, subclonal reconstructions, transcript abundance, splice calls and other core data generated by the ICGC/TCGA Pan-cancer Analysis of Whole Genomes Consortium is described in the marker paper[4] and available for download at https://dcc.icgc.org/releases/PCAWG. Additional information on accessing the data, including raw read files, can be found at https://docs.icgc.org/pcawg/data/. In accordance with the data access policies of the ICGC and TCGA projects, most molecular, clinical and specimen data are in an open tier which does not require access approval. To access potentially identifiable information, such as germline alleles and underlying sequencing data, researchers will need to apply to the TCGA Data Access Committee (DAC) via dbGaP (https://dbgap.ncbi.nlm.nih.gov/aa/wga.cgi?page = login) for access to the TCGA portion of the dataset, and to the ICGC Data Access Compliance Office (DACO; http://icgc.org/daco) for the ICGC portion. In addition, to access somatic single nucleotide variants derived from TCGA donors, researchers will also need to obtain dbGaP authorization. Derived PCAWG datasets described specifically in this manuscript can be found at the locations listed below. Additional relevant datasets are listed in the PCAWG study of pathways and networks[14].

Label	Synapse ID	ICGC DCC URL	ICGC DCC Filename	Access (Open/Controlled)
PCAWG driver *P*-values	syn8494939	https://dcc.icgc.org/releases/PCAWG/networks/	final_integration_results_2017_03_16.tar.gz	Open
PCAWG pathway and network method results	syn21413360	https://dcc.icgc.org/releases/PCAWG/networks/	pathway_and_network_method_results.tar.gz	Open
PCAWG pathway and network consensus results	syn11654843	https://dcc.icgc.org/releases/PCAWG/networks/	method_results_2017_10_10.tar.gz	Open
Enhancer-gene mappings	syn7201027	https://dcc.icgc.org/releases/PCAWG/networks/	map.enhancer.gene.txt.gz	Open
Coding and non-coding elements	syn21416282	https://dcc.icgc.org/releases/PCAWG/networks/	gene-coding-and-non-coding-elements.tar.gz	Open

Code availability

ActivePathways is freely available as an R package and source code on the GitHub repository https://github.com/reimandlab/ActivePathways and in the Comprehensive R Archive Network (CRAN). Additional scripts are available upon request. The core computational pipelines used by the PCAWG Consortium for alignment, quality control and variant calling are available to the public at https://dockstore.org/search?search=pcawg under the GNU General Public License v3.0, which allows for reuse and distribution.

References

1. Reimand, J. et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, cytoscape and enrichmentmap. *Nat. Protoc.* **14**, 482–517 (2019).
2. Weinstein, J. N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
3. Hudson, T. J. et al. International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
4. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1969-6 (2020).
5. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
6. Sinha, A. et al. The proteogenomic landscape of curable prostate cancer. *Cancer Cell* **35**, 414–427 e416 (2019).
7. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
8. Mi, H., Muruganujan, A. & Thomas, P. D. PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res.* **41**, D377–D386 (2013).
9. Kaimal, V., Bardes, E. E., Tabar, S. C., Jegga, A. G. & Aronow, B. J. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Res.* **38**, W96–W102 (2010).
10. Reimand, J., Kull, M., Peterson, H., Hansen, J. & Vilo, J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids Res.* **35**, W193–W200 (2007).
11. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
12. Huang, F. W. et al. Highly recurrent TERT promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
13. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* https://doi.org/10.1038/s41586-020-1965-x (2020).
14. Reyna, M. A. et al. Pathway and network analysis of more than 2500 whole cancer genomes. *Nat. Commun.* https://doi.org/10.1038/s41467-020-14367-0 (2020).
15. Curtis, C. et al. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
16. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
17. Fabregat, A. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
18. Brown, M. B. A method for combining non-independent, one-sided tests of significance. *Biometrics* **31**, 987–992 (1975).
19. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
20. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5**, e13984 (2010).
21. Quon, G. et al. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Med.* **5**, 29 (2013).
22. Fox, N. S., Haider, S., Harris, A. L. & Boutros, B. C. Landscape of transcriptomic interactions between breast cancer and its microenvironment. *Nat. Commun.* **10**, 3116 (2019).
23. Adams, S. et al. Prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancers from two phase III randomized adjuvant breast cancer trials: ECOG 2197 and ECOG 1199. *J. Clin. Oncol.* **32**, 2959–2966 (2014).
24. Sabatier, R. et al. A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res. Treat.* **126**, 407–420 (2011).
25. Wang, H. Y., Cheng, Z. & Malbon, C. C. Overexpression of mitogen-activated protein kinase phosphatases MKP1, MKP2 in human breast cancer. *Cancer Lett.* **191**, 229–237 (2003).
26. Carpenter, R. L. & Lo, H. W. Regulation of apoptosis by HER2 in breast cancer. *J. Carcinog. Mutagen.* https://doi.org/10.4172/2157-2518.S7-003 (2013).
27. Denu, R. A. et al. Centrosome amplification induces high grade features and is prognostic of worse outcomes in breast cancer. *BMC Cancer* **16**, 47 (2016).
28. Belin, S. et al. Dysregulation of ribosome biogenesis and translational capacity is associated with tumor progression of human breast cancer cells. *PLoS ONE* **4**, e7147 (2009).
29. Guimaraes, J. C. & Zavolan, M. Patterns of ribosomal protein expression specify normal and malignant human cells. *Genome Biol.* **17**, 236 (2016).
30. Varelas, X. The Hippo pathway effectors TAZ and YAP in development, homeostasis and disease. *Development* **141**, 1614–1626 (2014).
31. Yu, F. X., Zhao, B. & Guan, K. L. Hippo pathway in organ size control, tissue homeostasis, and cancer. *Cell* **163**, 811–828 (2015).
32. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
33. Galli, G. G. et al. YAP drives growth by controlling transcriptional pause release from dynamic enhancers. *Mol. Cell* **60**, 328–337 (2015).
34. Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).
35. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
36. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
37. Loh, Y. H. et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38**, 431–440 (2006).
38. Mo, J. S., Park, H. W. & Guan, K. L. The Hippo signaling pathway in stem cell biology and cancer. *EMBO Rep.* **15**, 642–656 (2014).
39. Leiserson, M. D. et al. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
40. Reyna, M. A., Leiserson, M. D. M. & Raphael, B. J. Identifying hierarchies of altered subnetworks. *Bioinformatics* **34**, i972–i980 (2018).
41. Pulido-Tamayo, S., Weytjens, B., De Maeyer, D. & Marchal, K. SSA-ME Detection of cancer driver genes using mutual exclusivity by small subnetwork analysis. *Sci. Rep.* **6**, 36257 (2016).
42. Verbeke, L. P. et al. Pathway relevance ranking for tumor samples through network-based data integration. *PLoS ONE* **10**, e0133503 (2015).
43. Kahraman, A. & von Mering, C. Pathogenic impact of isoform switches in 1209 cancer samples covering 27 cancer types using an isoform-specific interaction network. *bioRxiv* https://doi.org/10.1101/742379 (2019).
44. Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A. & Lopez-Bigas, N. Nucleotide excision repair is impaired by binding of transcription factors to DNA. *Nature* **532**, 264–267 (2016).
45. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
46. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
47. Wadi, L., Meyer, M., Weiser, J., Stein, L. D. & Reimand, J. Impact of outdated gene annotations on pathway enrichment analysis. *Nat. Methods* **13**, 705–706 (2016).
48. Liberzon, A. et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
49. Vaske, C. J. et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
50. Warde-Farley, D. et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **38**, W214–W220 (2010).
51. Cline, M. S. et al. Integration of biological networks and gene expression data using cytoscape. *Nat. Protoc.* **2**, 2366–2382 (2007).
52. Parker, J. S. et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **27**, 1160–1167 (2009).

Author contributions

J.B., H.Z., J.R. developed the method and the software package. M.P., H.Z., N.S., N.S.F., D.A.R., M.W.M., J.R. analysed the data. P.C.B. contributed to study design and supervision. J.R. designed and supervised the study. Members of the PCAWG Drivers and

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-019-13983-9.

Correspondence and requests for materials should be addressed to J.R.

Peer review information *Nature Communications* thanks Zhaohui Qin and other anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

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

PCAWG Drivers and Functional Interpretation Working Group

Federico Abascal[9], Samirkumar B. Amin[10,11,135], Gary D. Bader[12], Rameen Beroukhim[13,14,15], Johanna Bertl[16,17], Keith A. Boroevich[18,19], Søren Brunak[20,21], Peter J. Campbell[9,22], Joana Carlevaro-Fita[23,24,25], Dimple Chakravarty[26], Calvin Wing Yiu Chan[27,28], Ken Chen[29], Jung Kyoon Choi[30], Jordi Deu-Pons[31,32], Priyanka Dhingra[33,34], Klev Diamanti[35], Lars Feuerbach[36], J. Lynn Fink[37,38], Nuno A. Fonseca[39,40], Joan Frigola[31], Carlo Gambacorti-Passerini[41], Dale W. Garsed[42,43], Mark Gerstein[44,45,46,47], Gad Getz[13,15,48,49], Abel Gonzalez-Perez[32,50,51], Qianyun Guo[52], Ivo G. Gut[53,54], David Haan[55], Mark P. Hamilton[56], Nicholas J. Haradhvala[13,57], Arif O. Harmanci[47,58], Mohamed Helmy[59], Carl Herrmann[27,60,61], Julian M. Hess[13,62], Asger Hobolth[16,52], Ermin Hodzic[63], Chen Hong[28,36], Henrik Hornshøj[17], Keren Isaev[1,2], Jose M.G. Izarzugaza[20], Rory Johnson[24,64], Todd A. Johnson[18], Malene Juul[17], Randi Istrup Juul[17], Andre Kahles[65,66,67,68,69], Abdullah Kahraman[70,71,72], Manolis Kellis[13,73], Ekta Khurana[33,34,74,75], Jaegil Kim[13], Jong K. Kim[76], Youngwook Kim[77,78], Jan Komorowski[35,79], Jan O. Korbel[40,80], Sushant Kumar[46,47], Andrés Lanzós[24,25,64], Michael S. Lawrence[13,18,57], Donghoon Lee[47], Kjong-Van Lehmann[65,66,67,68,69], Shantao Li[47], Xiaotong Li[47], Ziao Lin[13,81], Eric Minwei Liu[33,34,82], Lucas Lochovsky[83,84,85,86], Shaoke Lou[46,47], Tobias Madsen[17], Kathleen Marchal[87,88], Iñigo Martincorena[9], Alexander Martinez-Fundichely[33,34,74], Yosef E. Maruvka[13,57,62], Patrick D. McGillivray[46], William Meyerson[47,89], Ferran Muiños[32,51], Loris Mularoni[32,51], Hidewaki Nakagawa[19], Morten Muhlig Nielsen[17], Keunchil Park[90,91], Kiejung Park[92], Jakob Skou Pedersen[17,52], Oriol Pich[32,51], Tirso Pons[93], Sergio Pulido-Tamayo[87,88], Benjamin J. Raphael[44], Iker Reyes-Salazar[51], Matthew A. Reyna[44], Esther Rheinbay[13,15,57], Mark A. Rubin[64,75,94,95,96], Carlota Rubio-Perez[32,51,97], Radhakrishnan Sabarinathan[32,51,98], S. Cenk Sahinalp[63,99,100], Gordon Saksena[13], Leonidas Salichos[46,47], Chris Sander[65,101,102,103], Steven E. Schumacher[13,104], Mark Shackleton[43,105], Ofer Shapira[13,102], Ciyue Shen[103,106], Raunak Shrestha[100], Shimin Shuai[1,12], Nikos Sidiropoulos[107], Lina Sieverling[28,36], Nasa Sinnott-Armstrong[13,108], Lincoln D. Stein[1,12], Joshua M. Stuart[55], David Tamborero[32,51], Grace Tiao[13], Tatsuhiko Tsunoda[18,109,110,111], Husen M. Umer[35,112], Liis Uusküla-Reimand[113,114], Alfonso Valencia[115,116], Miguel Vazquez[115,117], Lieven P.C. Verbeke[88,118], Claes Wadelius[119], Lina Wadi[1], Jiayin Wang[120,121,122], Jonathan Warrell[46,47], Sebastian M. Waszak[80], Joachim Weischenfeldt[80,107,123], David A. Wheeler[124,125], Guanming Wu[126], Jun Yu[127,128], Jing Zhang[47], Xuanping Zhang[120,129], Yan Zhang[47,130,131], Zhongming Zhao[132], Lihua Zou[133] & Christian von Mering[72,134]

[9]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [10]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [11]Quantitative & Computational Biosciences Graduate Program, Baylor College of Medicine, Houston, TX 77030, USA. [12]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada. [13]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [14]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA. [15]Harvard Medical School, Boston, MA 02115, USA. [16]Department of Mathematics, Aarhus University, 8000 Aarhus, Denmark. [17]Department of Molecular Medicine (MOMA), Aarhus University Hospital, 8200 Aarhus N, Denmark. [18]Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. [19]RIKEN Center for Integrative Medical Sciences, Yokohama, Kanagawa 230-0045, Japan. [20]Technical University of Denmark, 2800 Lyngby, Denmark. [21]University of Copenhagen, 2200 Copenhagen, Denmark. [22]Department of Haematology, University of Cambridge, Cambridge CB2 2XY, UK. [23]Department for BioMedical Research, University of Bern, 3008 Bern, Switzerland. [24]Department of Medical Oncology, Inselspital, University Hospital and University of Bern, 3010 Bern, Switzerland. [25]Graduate School for Cellular and Biomedical Sciences, University of Bern, 3012 Bern, Switzerland. [26]Department of Genitourinary Medical Oncology—Research, Division of Cancer Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [27]Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [28]Faculty of Biosciences, Heidelberg University, 69120 Heidelberg, Germany. [29]University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA. [30]Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea. [31]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 8003 Barcelona, Spain. [32]Research Program on Biomedical Informatics, Universitat Pompeu Fabra, 08002 Barcelona, Spain. [33]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY 10065, USA. [34]Institute for Computational Biomedicine, Weill Cornell Medicine, New York, NY 10021, USA. [35]Science for Life Laboratory, Department of Cell and Molecular Biology, Uppsala University, SE-75124 Uppsala, Sweden. [36]Division of Applied Bioinformatics, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany. [37]Barcelona Supercomputing Center, 08034 Barcelona, Spain. [38]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, St Lucia, QLD 4072, Australia. [39]CIBIO/InBIO—Research Center in Biodiversity and Genetic Resources, Universidade do Porto, 4485-601 Vairão, Portugal. [40]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. [41]University of Milano Bicocca, 20052 Monza, Italy. [42]Peter MacCallum Cancer Centre, Melbourne, VIC 3000, Australia. [43]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, VIC 3052, Australia. [44]Department of Computer Science, Princeton University, Princeton, NJ 08540, USA. [45]Department of Computer Science, Yale University, New Haven, CT 06520, USA. [46]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [47]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06520, USA. [48]Center for Cancer Research, Massachusetts General Hospital, Boston, MA 02129, USA. [49]Department of Pathology, Massachusetts General Hospital, Boston, MA 02115, USA. [50]Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), 08003 Barcelona, Spain. [51]Institute for Research in Biomedicine (IRB Barcelona), 08028 Barcelona, Spain. [52]Bioinformatics Research Centre (BiRC), Aarhus University, 8000 Aarhus, Denmark. [53]CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), 08028 Barcelona, Spain. [54]Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain. [55]Biomolecular Engineering Department, University of California, Santa Cruz, Santa Cruz, CA 95064, USA. [56]Department of Internal Medicine, Stanford University, Stanford, CA 94305, USA. [57]Massachusetts General Hospital, Boston, MA 02114, USA. [58]Center for Precision Health, School of Biomedical Informatics, University of Texas Health Science Center, Houston, TX 77030, USA. [59]The Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada. [60]Health Data Science Unit, University Clinics, 69120 Heidelberg, Germany. [61]Institute of Pharmacy and Molecular Biotechnology and BioQuant, Heidelberg University, 69120 Heidelberg, Germany. [62]Massachusetts General Hospital Center for Cancer Research, Charlestown, MA 02129, USA. [63]Simon Fraser University, Burnaby, BC V5A 1S6, Canada. [64]Department for Biomedical Research, University of Bern, 3008 Bern, Switzerland. [65]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [66]Department of Biology, ETH Zurich, Wolfgang-Pauli-Strasse 27, 8093 Zürich, Switzerland. [67]Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland. [68]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland. [69]University Hospital Zurich, 8091 Zurich, Switzerland. [70]Clinical Bioinformatics, Swiss Institute of Bioinformatics, 1202 Geneva, Switzerland. [71]Institute for Pathology and Molecular Pathology, University Hospital Zurich, 8091 Zurich, Switzerland. [72]Institute of Molecular Life Sciences, University of Zurich, 8057 Zurich, Switzerland. [73]MIT Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [74]Englander Institute for Precision Medicine, Weill Cornell Medicine, New York, NY 10065, USA. [75]Meyer Cancer Center, Weill Cornell Medicine, New York, NY 10065, USA. [76]Research Core Center, National Cancer Centre Korea, Goyang-si 410-769, South Korea. [77]Department of Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [78]Samsung Genome Institute, Seoul 06351, South Korea. [79]Institute of Computer Science, Polish Academy of Sciences, 01-248 Warsawa, Poland. [80]Genome Biology Unit, European Bioinformatics Institute (EMBL-EBI), 69117 Heidelberg, Germany. [81]Harvard University, Cambridge, MA 02138, USA. [82]Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA. [83]Department of Molecular Biophysics and Biochemistry, New Haven, CT 06520, USA. [84]Program in Computational Biology and Bioinformatics, New Haven, CT 06520, USA. [85]The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA. [86]Yale University, New Haven, CT 06520, USA. [87]Department of Information Technology, Ghent University, 9000 Ghent, Belgium. [88]Department of Plant Biotechnology and Bioinformatics, Ghent University, 9000 Ghent, Belgium. [89]Yale School of Medicine, Yale University, New Haven, CT 06520, USA. [90]Division of Hematology-Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [91]Samsung Advanced Institute for Health Sciences and Technology, Sungkyunkwan University School of Medicine, Seoul 06351, South Korea. [92]Cheonan Industry-Academic Collaboration Foundation, Sangmyung University, Cheonan 31066, South Korea. [93]Spanish National Cancer Research Centre, 28029 Madrid, Spain. [94]Bern Center for Precision Medicine, University Hospital of Bern, University of Bern, 3008 Bern, Switzerland. [95]Englander Institute for Precision Medicine, Weill Cornell Medicine and NewYork Presbyterian Hospital, New York, NY 10021, USA. [96]Pathology and Laboratory, Weill Cornell Medical College, New York, NY 10021, USA. [97]Vall d'Hebron Institute of Oncology: VHIO, 08035 Barcelona, Spain. [98]National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore 560065, India. [99]Indiana University, Bloomington, IN 47405, USA. [100]Vancouver Prostate Centre, Vancouver, BC V6H 3Z6, Canada. [101]cBio Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA. [102]Dana-Farber Cancer Institute, Boston, MA 02215, USA. [103]Department of Cell Biology, Harvard Medical School, Boston, MA 02115, USA. [104]Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [105]Peter MacCallum Cancer Centre and University of Melbourne, Melbourne, VIC 3000, Australia. [106]cBio Center, Dana-Farber Cancer Institute, Boston, MA 02215, USA. [107]Finsen Laboratory and Biotech Research & Innovation Centre (BRIC), University of Copenhagen, 2200 Copenhagen, Denmark. [108]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA. [109]CREST, Japan Science and Technology Agency, Tokyo 113-0033, Japan. [110]Department of Medical Science Mathematics, Tokyo Medical and Dental University, Bunkyo-ku, Tokyo 113-8510, Japan. [111]Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Bunkyo-ku, Tokyo 113-0033, Japan. [112]Science for Life Laboratory, Department of Oncology-Pathology, ,Karolinska Institutet, SE-17121 Stockholm, Sweden. [113]Department of Gene Technology, Tallinn University of Technology, 12616 Tallinn, Estonia. [114]Genetics & Genome Biology Program, SickKids Research Institute, The

Hospital for Sick Children, Toronto, ON M5G 1X8, Canada. [115]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain. [116]Institució Catalana de Recerca i Estudis Avançats (ICREA), 08010 Barcelona, Spain. [117]Department of Clinical and Molecular Medicine, Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology, 7030 Trondheim, Norway. [118]Department of Information Technology, Ghent University, Interuniversitair Micro-Electronica Centrum (IMEC), 9000 Ghent, Belgium. [119]Science for Life Laboratory, Department of Immunology, Genetics and Pathology, Uppsala University, SE-75108 Uppsala, Sweden. [120]School of Computer Science and Technology, Xi'an Jiaotong University, 710048 Xi'an, China. [121]School of Electronic and Information Engineering, Xi'an Jiaotong University, 710048 Xi'an, China. [122]The McDonnell Genome Institute at Washington University, St Louis, MO 63108, USA. [123]Department of Urology, Charité Universitätsmedizin Berlin, 10117 Berlin, Germany. [124]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030, USA. [125]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA. [126]Oregon Health & Sciences University, Portland, OR 97239, USA. [127]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong, China. [128]Second Military Medical University, Shanghai 200433, China. [129]The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [130]Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA. [131]The Ohio State University Comprehensive Cancer Center (OSUCCC – James), Columbus, OH 43210, USA. [132]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. [133]Department of Biochemistry and Molecular Genetics, Feinberg School of Medicine, Northwestern University, Chicago, IL 60637, USA. [134]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland. [135]Present address: The Jackson Laboratory for Genomic Medicine, Farmington, CT 06032, USA