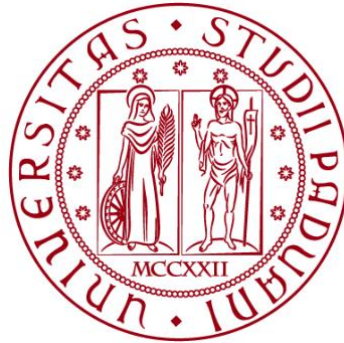


**UNIVERSITÀ DEGLI STUDI DI PADOVA**

**DIPARTIMENTO DI BIOLOGIA**

**Corso di Laurea in Biologia Molecolare**



**ELABORATO DI LAUREA**

**Analisi esplorativa dei profili di instabilità  
genomica in tumore**

**Tutor: Prof.ssa Chiara Romualdi  
Dipartimento di Biologia - DiBio**

**Laureando: Alberto Lupatin**

**ANNO ACCADEMICO 2022/2023**





# INDICE

ABSTRACT .....	1
1. STATO DELL'ARTE: Introduzione al Problema Biologico .....	3
1.1. Introduzione .....	3
1.2. Il ruolo delle mutazioni somatiche nello sviluppo del cancro .....	3
1.3. L'identificazione di Pattern nella tipizzazione dei tumori .....	4
1.4. Applicazioni medico-biologiche delle signatures di CNV .....	6
2. APPROCCIO SPERIMENTALE: Descrizione delle Metodologie .....	7
2.1. Caratterizzazione delle signatures .....	7
2.2. Generazione dei dataset .....	8
2.3. Estrazione delle signatures .....	8
2.3.1. Prima elaborazione dei dati grezzi .....	8
2.3.2. SigProfilerExtractor .....	9
2.4. Etichettatura e Comparazione delle signatures .....	11
2.5. Analisi di Sopravvivenza .....	11
3. RISULTATI: Analisi critica dell'approccio sperimentale .....	13
3.1. Selezione delle signatures estratte .....	13
3.2. Matrice di signatures .....	14
3.3. Confronto signatures PCAWG e TCGA .....	15
3.4. Distribuzione delle signatures pan-cancer e correlazione con i processi mutagenici .....	17
3.5. Analisi cliniche .....	19
4. DISCUSSIONE .....	21
5. BIBLIOGRAFIA .....	22
6. APPENDICE 1: Esperienza di Stage .....	23
6.1. Studi analizzati .....	23
6.2. Metodi utilizzati .....	23
6.3. Risultati ottenuti .....	24
6.3.1. Correlazioni .....	24
6.3.2. Heatmap .....	25
6.3.3. Analisi Cliniche .....	26
7. APPENDICE 2: Grafici Supplementari .....	27



## ABSTRACT

Le mutazioni somatiche rappresentano la causa principale nello sviluppo del cancro. I pattern di tali mutazioni rappresentano degli indicatori fondamentali nella mutagenesi e cancerogenesi e sono definite a livello generico come instabilità cromosomica (CIN).

CIN ha conseguenze importanti a livello genomico tra cui: la perdita o l'aumento del numero dei geni *driver*, riarrangiamenti, formazione di micronuclei e l'attivazione della risposta immunitaria. Ciò può essere associato allo stadio della malattia, metastasi e una ridotta reazione alla terapia.

Le cause sono riconducibili sia a mutazioni su bassa scala (mutazioni puntiformi, inserzioni e delezioni di piccoli frammenti) sia a mutazioni su larga scala, dette variazioni strutturali (SV). In particolare, l'alterazione del numero di ripetizioni (CNA) rappresenta il caso più frequente di SV.

Tipi di mutazione diversi possono essere associate a caratteristiche cliniche ed eziologiche comuni tra loro e possono essere quindi raggruppate in una categoria, detta *signature* di instabilità. Tali *signature* sono tutt'oggi oggetto di analisi genomiche, in quanto permettono di ottimizzare la terapia e di predire la comparsa di una recidiva.



# 1. STATO DELL'ARTE: *Introduzione al Problema Biologico*

## 1.1. Introduzione

In questo elaborato verranno presentati i principali risultati del lavoro di Tao *et al.* dal titolo: “*The repertoire of copy number alteration signatures in human cancer*”.

Lo studio esplora la tematica delle *signatures* di copy number, a partire dall'estrazione delle stesse da due grandi dataset di pazienti (PCAWG e TCGA) per poi proseguire sulla loro utilità in ambito clinico e concludendo gettando le basi per futuri approfondimenti sulle *signatures* estratte.

L'elaborato è diviso in più sezioni.

Nella prima verrà introdotto l'argomento delle mutazioni somatiche e delle *signatures*, con un focus sulla loro utilità nell'ambito della ricerca contro il cancro.

Successivamente verranno presentati i metodi utilizzati dagli autori dell'articolo e i rispettivi risultati ottenuti, concludendo con la discussione degli stessi.

Infine, sono presenti due appendici riguardanti l'esperienza di tirocinio svoltasi presso il laboratorio della Professoressa Romualdi. In particolare, nella prima appendice si troverà la descrizione del lavoro eseguito mentre nella seconda si troveranno dei grafici supplementari.

## 1.2. Il ruolo delle mutazioni somatiche nello sviluppo del cancro

Il genoma di ogni cellula dell'organismo subisce costantemente una pletora di processi, endogeni ed esogeni, che possono portare ad un danno al DNA. Nella maggior parte dei casi questo danno viene riparato tramite gli appositi meccanismi di riparazione del genoma e rimane confinato nella cellula da cui si è originato. Tuttavia, se i meccanismi di riparazione non riconoscono o non riescono a riparare tale danno, il DNA conterrà una mutazione che verrà mantenuta nella cellula stessa e verrà trasmessa alle cellule figlie. Questo tipo di mutazioni vengono definite somatiche.

Le mutazioni somatiche rappresentano la causa principale nello sviluppo del cancro. Tali mutazioni portano ad un cambiamento della sequenza nucleotidica del DNA a causa di eventi come le sostituzioni di un singolo (SBS – Single Base Substitution) o due nucleotidi (DBS – Double Base Substitutions) o inserzioni e delezioni (INDEL).

Le mutazioni somatiche possono anche riguardare milioni di paia di basi e



coinvolgere più cromosomi. In tal caso ci si riferisce a varianti strutturali (SVs – Structural Variants), nei quali i cromosomi vengono fusi parzialmente, o a variazioni del numero di copie di un segmento genico lungo non meno di 1 kb (CNV – Copy Number Variations).

Le CNV possono essere causate da svariati meccanismi, ma possono essere suddivise principalmente in due categorie: duplicazione del numero di copie dell'intero genoma (WGD – Whole-Genome Duplication) o perdita di eterozigosi (LOH – Loss Of Heterozygosity).

Le mutazioni sopracitate riguardano principalmente i geni definiti *driver* ovvero geni che, se attivi, permettono la selezione positiva delle cellule tumorali. Mutazioni di questo tipo hanno un impatto significativo nello sviluppo della malattia, in particolare contribuiscono all'espansione delle cellule tumorali e alla loro capacità di invadere tessuti circostanti. Esiste tuttavia un altro tipo di mutazioni, ovvero le mutazioni “passeggere”, meno studiate rispetto alle *driver* in quanto avvengono casualmente e non contribuiscono in modo significativo allo sviluppo del cancro. Nonostante ciò, esse contengono informazioni sui processi mutagenici avvenuti nel corso della vita della cellula e delle cellule da cui deriva.

### 1.3. L'identificazione di Pattern nella tipizzazione dei tumori

Tramite l'analisi bioinformatica delle sequenze di DNA ottenute tramite sequenziamento next-gen di numerosi campioni tumorali, è stato possibile identificare dei pattern di alterazioni genomiche. In particolare, sono stati individuati dei pattern genomici che sono rappresentativi del tipo di mutazione responsabile di quel particolare tipo di cancro. Tali pattern sono definiti *signatures* tumorali.

Le *signatures* rappresentano quindi il risultato di numerosi eventi mutagenici che hanno portato alla cancerogenesi, come rappresentato in *Figura 1*.



maggior parte dei geni analizzati, indicando come le CNV abbiano un effetto diretto sulla trascrizione genica.

Inoltre, si possono ricavare informazioni più precise riguardo i processi mutagenici che sono avvenuti e che hanno portato alla formazione della cellula cancerogena [2].

Tuttavia, i pathway precisi che portano alla cancerogenesi rimangono ancora da definire.

#### **1.4. Applicazioni medico-biologiche delle *signatures* di CNV**

Dal punto di vista applicativo, le *signatures* sono fondamentali nel campo della ricerca e della terapia contro il cancro.

Nella ricerca rappresentano un importante tool di analisi per identificare indirettamente processi di mutazione, riparazione, replicazione del DNA e per identificare gli enzimi coinvolti in questi meccanismi.

Per quanto riguarda la terapia, le *signatures* si sono rivelate fondamentali per predire le conseguenze della terapia oncologica nell'organismo e per ottimizzare la stessa.

Sono ormai note le conseguenze *long-term* dei farmaci antitumorali, dovute principalmente alle loro capacità mutageniche e citotossiche. Tuttavia, solo di recente sono state analizzate le *signatures* di mutazioni causate da essi, permettendo di correlare l'insorgenza di tumori secondari, o metastasi, e l'utilizzo di immuno o chemioterapie per il trattamento del tumore primario[3]. Infine, le *signatures* possono essere utilizzate, a livello clinico, come biomarker predittivi dell'efficacia di determinati farmaci, in modo da ottimizzare la terapia in base al tipo di cancro diagnosticato.

## 2. APPROCCIO SPERIMENTALE: *Descrizione delle Metodologie*

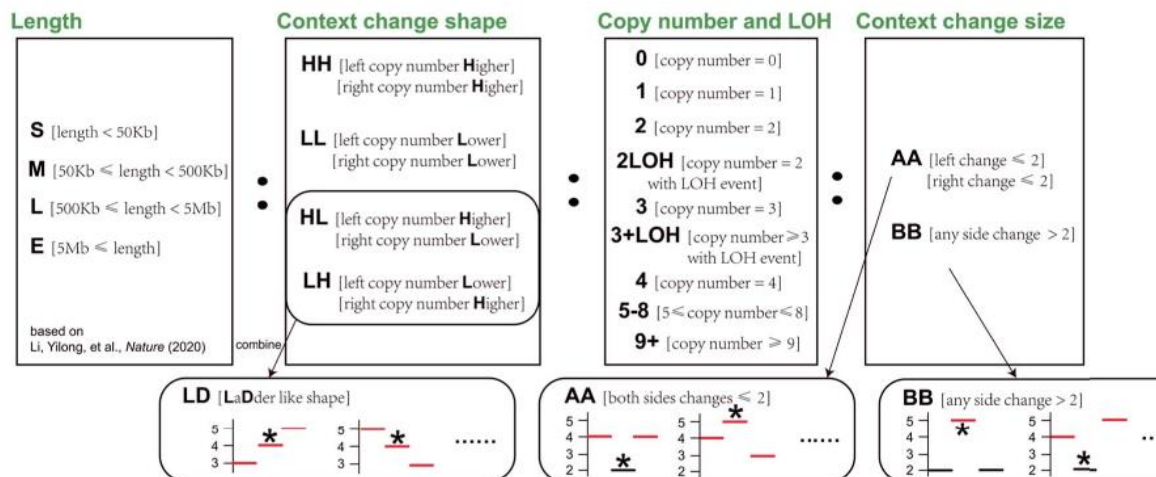
### 2.1. Caratterizzazione delle *signatures*

Uno step fondamentale che precede l'analisi delle *signatures* è la classificazione delle stesse.

Nel lavoro di *Tao et. al* per ogni segmento che presenta alterazioni nel copy number, sono state prese in considerazione 176 elementi, riassumibili nelle seguenti caratteristiche (*Figura 2*):

1. Segmento normale (CN = 2), amplificato (CN > 2) o deleto (CN < 2)
2. Dimensione: small (S – lunghezza < 50 kb); medium (M – 50 kb ≤ lunghezza < 500 kb); large (L – 500 kb ≤ lunghezza < 5 Mb); extreme large (E – 5 Mb ≤ lunghezza).
3. “Contesto” del segmento: verifica se è presente un cambiamento di forma, ovvero un numero di copie maggiore o minore a destra o a sinistra del segmento target.
4. Numero di copie in valore assoluto associato alla perdita o meno di eterozigosi (LOH – Loss Of Heterozygosity).

Sono state identificate 7 categorie: 0, 1, 2, 2 con LOH, 3, ≥3 con LOH, 4, da 5 a 8 e ≥9.



*Figura 2: criteri per la classificazione delle CNA*

## 2.2. Generazione dei dataset

Sono stati utilizzati due dataset, uno dal PCAWG (Pan Cancer Analysis Whole Genome) e uno dall'Atlas (TCGA) .

Il dataset PCAWG è stato scaricato da UCSC Xena e contiene le sequenze complete (WGS - Whole Genome Sequence) di 2778 campioni relativi a 32 tipi di cancro diversi. Sono stati combinati i dati provenienti da 6 algoritmi per l'identificazione del numero di copie: ABSOLUTE, ACEseq, Battenberg, CloneHD, JaBbA e Sclust. Il dataset TCGA è composto da 10851 campioni relativi a 33 tipi di cancro diversi generati tramite l'array Affymetrix Genome-Wide Human SNP 6.0 (SNP6) utilizzando il workflow ASCAT2.

I valori di CN sia di PCAWG che TCGA sono interi.

## 2.3. Estrazione delle *signatures*

L'estrazione delle *signatures* è stata eseguita tramite il tool *SigProfilerExtractor* v1.0.17, sviluppato da Alexandrov *et al.* [4]. Esso rappresenta il gold-standard per l'estrazione *de novo* delle *signatures* per tutti i tipi di mutazione somatica a partire da una matrice di dati. In particolare, la funzione identifica: il numero di *signatures* operative, le loro attività in ogni campione e calcola la probabilità per ogni *signature* di causare un tipo di mutazione specifica del campione tumorale. Questo tool è stato successivamente integrato in *Sigminer*, un pacchetto CRAN sviluppato da Tao *et. al.*

La procedura di identificazione delle *signatures* prevede quindi due step: generazione della matrice con i tipi di mutazione per campione ed estrazione effettiva delle *signatures*.

### 2.3.1. Prima elaborazione dei dati grezzi

Dopo aver scaricato i profili ASCAT di CN dalla pagina di [Github](#), è necessario unirli in un'unica matrice.

I singoli file di testo scaricati sono in formato ASCAT e contengono informazioni riguardo il barcode del campione, il cromosoma nel quale è presente la mutazione, la posizione di inizio e fine della mutazione e il numero di copie.

Il file generato avrà il formato .TSV.

Successivamente si utilizza un ulteriore tool fornito da Alexandrov *et. al.*, *SigProfilerMatrixGenerator*, necessario per categorizzare le mutazioni in base alla loro tipologia: SBS, DBS e brevi INDEL.

L'utilizzo di questo tool è necessario per ridurre la mole di dati della matrice generata precedentemente e per identificare più efficacemente le possibili *signatures* nel genoma.

Dopo aver installato *SigProfilerMatrixGenerator*, è necessario ricavare i files del genoma di riferimento. Ciò può essere eseguito tramite la funzione preposta presente nel pacchetto. Nel lavoro di Tao *et. al* viene utilizzato il genoma GRCh37 (Genome Reference Consortium Human Reference 37), rilasciato ad aprile 2011, aggiornato a settembre 2023 e scaricato dal database ENSEMBL (versione 93.37).

La funzione riceve in input i seguenti parametri:

1. La posizione nel computer del file generato inizialmente
2. Il formato del file (ASCAT)
3. La posizione nel computer della cartella generata in output
4. Il nome del progetto

Restituisce una matrice in cui le righe rappresentano i diversi tipi di mutazione, le colonne corrispondono ai campioni analizzati e i valori indicano quante volte è presente un tipo di mutazione in quel particolare sample.

### **2.3.2. *SigProfilerExtractor***

In seguito, si procede all'estrazione delle *signatures* tramite l'utilizzo di *SigProfilerExtractor*, un pacchetto Python che consente l'analisi di diverse tipologie di mutazioni. Questo strumento seleziona automaticamente il numero delle *signatures* e associa le *signatures* estratte *de novo* con quelle presenti nel database COSMIC.

Di default *SigProfilerExtractor* richiede solamente un parametro in input, ovvero la posizione del file contenente il dataset generato in precedenza tramite *SigProfilerMatrix Generator*.

Il tool procede con i seguenti 6 step: riduzione delle dimensioni della matrice, ricampionamento, utilizzo dell'algoritmo di fattorizzazione non-negativa (NMF - Non-negative Matrix Factorization), iterazione, clusterizzazione e controllo dei risultati.

La funzione assembla le matrici NMF per ridurre la mole di dati da calcolare, e ottimizzare quindi il processo.

Per ogni NMF vengono eseguite da 10,000 a 1mln di iterazioni, in modo da ottenere un risultato sufficientemente stabile. Questo step viene poi ripetuto 100 volte ricampionando i dati.

In seguito vengono categorizzate e raggruppate le matrici scomposte in modo da ottenere da 2 a 30 *signatures* attive.

Infine, le *signatures* estratte vengono confrontate con le *signatures* reference presenti in COSMIC. Le *signatures* sono indicate come nuove quando, dopo averle scomposte, non è stato possibile associarle a nessuna *signature* del COSMIC.

*SigProfilerExtractor* è stato applicato a entrambi i database, con un totale di 30 *signatures* estratte per PCAWG e per TCGA.

Il tool fornisce in output numerosi file che contengono matrici e grafici ottenuti dal confronto di diverse variabili, tra cui il tipo di mutazione, le *signatures* estratte e i campioni.

Tra tutti, il file utile per le analisi successive è indicato come *Activities* ed è una matrice le cui righe sono i campioni, le colonne sono le *signatures* e i valori corrispondono al numero di mutazioni (di tutti i tipi) per ogni *signature*. Tali valori vanno da 0 a 875 ma sono stati successivamente normalizzati in base al valore massimo per ogni *signature*, in modo da ottenere dei valori più facilmente analizzabili che vanno da 0 a 1.

In questo modo si riesce ad attribuire una probabilità di appartenenza di ogni *signature* per uno specifico campione tumorale.

Le *signatures* così ottenute sono state ulteriormente selezionate utilizzando la funzione *show\_sig\_number\_survey* presente all'interno del pacchetto *Sigminer*.

Tale funzione riordina e genera un grafico delle *signatures* estratte sulla base della stabilità e della media della cosine similarity. Le *signatures* vengono quindi disposte in ordine crescente di tali parametri e vengono selezionate manualmente quelle che presentano valori di cosine similarity.

Sono state selezionate 14 *signatures* dal dataset PCAWG e 20 *signatures* dai campioni TCGA.

#### **2.4. Etichettatura e Comparazione delle *signatures***

È stato eseguito un primo processo di denominazione delle *signatures* in base alla loro attività su tutti i campioni.

Sono state quindi nominate da CNS1 a CNS14 le *signatures* dei dati di PCAWG e da Sig1 a Sig20 le *signatures* di TCGA.

È stato inoltre eseguito un secondo lavoro di etichettatura necessario per comparare meglio l'attività delle *signatures* tra i due datasets.

Sono stati aggiunti quindi altri *labels* sulla base di:

- Cosine similarity sulla somiglianza tra le *signatures*: alta (H - > 0.8), intermedia (I - tra 0.51 e 0.8) e nessun match (< 0.51)
- Match tra le *signatures* PCAWG e TCGA. Per esempio, Sig4 di TCGA è simile a CNS1 e CNS5 in maniera intermedia (M), quindi la *signature* prende il nome di: Sig3-CNS1(M)-CNS5(M)
- Uno step finale per sistemare le *signatures* di PCAWG che matchano con più *signatures* TCGA. In questo caso viene aggiunto a pedice il numero di *signatures* del TCGA che sono simili alla *signature* PCAWG di riferimento. Per esempio, Sig8-CNS5(M)\_3 indica che Sig8 è la terza *signature* simile a CNS5.

## 2.5. Analisi di Sopravvivenza

L'associazione tra l'attività delle *signatures* e la sopravvivenza del paziente è stata identificata applicando il modello a rischi proporzionali di *Cox* per ogni tipo di cancro.

Tale modello è stato eseguito in R tramite il pacchetto *ezcox*.

Sono stati successivamente confrontati gli effetti dell'attività delle *signatures* sulla prognosi nelle tipologie di cancro PCAWG e TCGA che dispongono di un numero sufficiente di pazienti (> 50) dei quali vi è la disponibilità di dati su profilo delle CNA e sulla sopravvivenza complessiva.

Per ogni modello *Cox* analizzato viene riportato uno *score Z*, un indice di significatività della correlazione tra la *signature* in esame e la sopravvivenza.

Se  $Z > 1.96$ , l'attività della *signature* è associata con un minor tempo di vita, se, al contrario,  $Z < -1.96$ , l'attività della *signature* è associata con un aumento della sopravvivenza.

Infine, a partire dai dati ottenuti sulla sopravvivenza è stato realizzato un grafico tramite lo stimatore di *Kaplan-Meier*, implementato in R con il pacchetto *survival*. Tale algoritmo permette di stimare la funzione di sopravvivenza dei dati relativi alla durata del *follow-up*.

Il valore di *cut-off* per questa analisi è stato determinato dalla funzione *surv\_cutpoint* presente nel pacchetto R *Survminer*.

In questo modo si ottengono grafici in cui sono presenti, per ogni *signature*, 2 curve: una relativa alla sopravvivenza (in giorni) quando essa è attiva e una relativa alla sopravvivenza quando non è attiva.



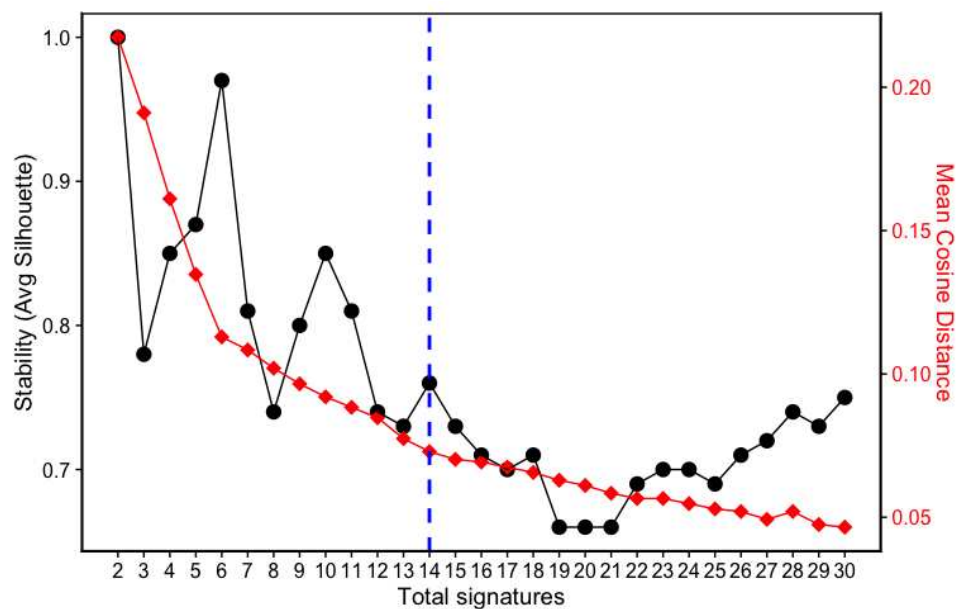


### 3. RISULTATI: *Analisi critica dell'approccio sperimentale*

#### 3.1. Selezione delle *signatures* estratte

Le *signatures* estratte da PCAWG e TCGA tramite il pacchetto *SigProfilerExtractor* sono state selezionate tramite la funzione *show\_sig\_number\_survey* presente in *Sigminer*, come descritto nei metodi.

In *Figura 3* è rappresentato il grafico relativo alle *signatures* PCAWG. Gli autori hanno deciso di tenere 14 *signatures* in quanto presentavano una stabilità relativamente alta e una distanza media bassa.

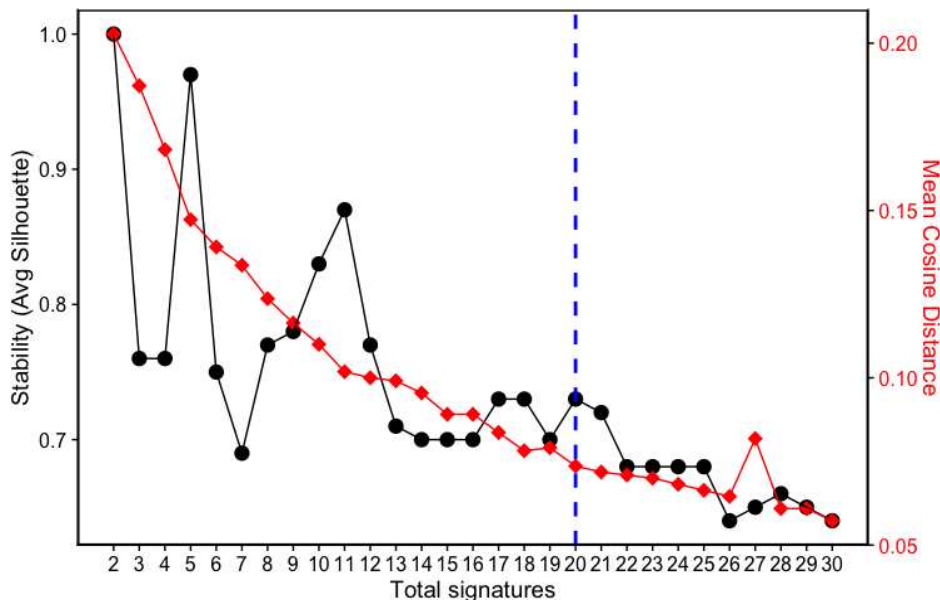


**Figura 3:** grafico relativo alla stabilità delle *signatures* PCAWG.

I punti neri indicano la stabilità di ogni *signature* mentre i rombi rossi indicano la media della *cosine distance*.

La linea tratteggiata blu indica il valore di *threshold* per la *cosine similarity*, posto dagli autori, al di sotto del quale hanno scartato le *signatures*.

In *Figura 4* è rappresentato il grafico relativo alle *signatures* TCGA. Gli autori hanno deciso di tenere 20 *signatures* per gli stessi motivi.



**Figura 4:** grafico relativo alla stabilità delle *signatures* TCGA.

I punti neri indicano la stabilità di ogni *signature* mentre i rombi rossi indicano la media della cosine distance.

La linea tratteggiata blu indica il valore di threshold per la cosine similarity, posto dagli autori, al di sotto del quale hanno scartato le *signatures*.

Il maggior numero di *signatures* estratte dal dataset TCGA, rispetto a PCAWG, è un'osservazione spiegabile dal maggior numero di campioni di partenza di TCGA (più di 10000) rispetto a PCAWG (2778).

### 3.2. Matrice di *signatures*

Una volta estratte le *signatures* rilevanti, si può costruire la matrice di *signatures* definitiva, la quale presenta nelle righe i campioni e nelle colonne le *signatures*.

I valori indicati rappresentano le probabilità che una determinata *signature* sia associata allo specifico campione. Tali valori sono compresi tra 0 e 1, dove 1 indica la certezza che la *signature* in questione sia coinvolta nello sviluppo della malattia. In entrambi i casi (PCAWG e TCGA) si tratta di matrici molto complesse, dato l'alto numero di campioni.

Nella pagina successiva è rappresentata una parte, per motivi di impaginazione, della matrice delle *signatures* PCAWG.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	sample	cancer_type	CNS1	CNS2	CNS3	CNS4	CNS5	CNS6	CNS7	CNS8	CNS9	CNS10	CNS11	CNS12	CNS13	CNS14
2	SP1003	Bladder-TCC	0,05976096	0,15936255	0,01992032	0,23904382	0,1314741	0	0,09163347	0,01195219	0,18326693	0,01992032	0,03187251	0,0438247	0	0,00796813
3	SP10084	Breast	0,10545455	0,10545455	0	0,38909091	0,26909091	0,00363636	0,01090909	0	0,06363636	0	0	0,00727273	0,03090909	0,01454545
4	SP1009	Bladder-TCC	0	0,03424658	0,08219178	0	0	0	0,10273973	0,32876712	0,03424658	0,28767123	0,03424658	0,0890411	0,00684932	0
5	SP10150	Breast	0,29090909	0,00909091	0,05454545	0	0,34545455	0,07272727	0,00909091	0	0,10909091	0	0	0,10909091	0	0
6	SP101515	Ovary-AdenoC	0,05263158	0,03827751	0,09090909	0	0,01435407	0,04784689	0,08133971	0,03827751	0,05263158	0,16267943	0,10526316	0,11004785	0	0,20574163
7	SP101519	Ovary-AdenoC	0,04929577	0,11971831	0,01408451	0,02112676	0,18309859	0,18309859	0	0,05633803	0,09859155	0,04929577	0,03521127	0,15492958	0	0,03521127
8	SP101521	Ovary-AdenoC	0,27150084	0,26138828	0,02023609	0,12310287	0,04215852	0,07082631	0,06576728	0	0,0623946	0	0,01686341	0,02866779	0,03709949	0
9	SP101523	Ovary-AdenoC	0	0,05511811	0	0,00787402	0	0,23228346	0,01574803	0	0,01181102	0,2519685	0,03149606	0	0,39370079	0
10	SP101526	Ovary-AdenoC	0,23320158	0,09288538	0,01778656	0,12648221	0,34980237	0,09288538	0,02766798	0	0,03162055	0,00988142	0,00592885	0,01185771	0	0
11	SP101528	Ovary-AdenoC	0,04379562	0,04379562	0,06569343	0,09489051	0,01459854	0	0,10218978	0	0,05109489	0,24087591	0,11678832	0,10948905	0,08029197	0,03649635
12	SP101532	Ovary-AdenoC	0	0,02926829	0,08292683	0	0,0097561	0,03414634	0,03902439	0,22439024	0,03414634	0,05853659	0,24878049	0,02926829	0	0,2097561
13	SP101536	Ovary-AdenoC	0,12096774	0	0,0483871	0,05645161	0,11290323	0,25806452	0,02419355	0,00806452	0,20967742	0	0,03225806	0,0483871	0,06451613	0,01612903
14	SP101540	Ovary-AdenoC	0,08783784	0,26689189	0,04391892	0,07094595	0,24324324	0,00675676	0,05743243	0,00675676	0,05067568	0,01013514	0,02364865	0,09121622	0,04054054	0
15	SP101544	Ovary-AdenoC	0,12704918	0,18032787	0,04098361	0,14754098	0,16803279	0	0,05327869	0	0,07786885	0,03278689	0	0,08196721	0,08606557	0,00409836
16	SP101548	Ovary-AdenoC	0,03496503	0	0,02797203	0	0,04895105	0	0,17482517	0,03496503	0,25874126	0,18881119	0,0979021	0,01398601	0,11888112	0
17	SP101552	Ovary-AdenoC	0,14945652	0,17663043	0,0326087	0	0,26902174	0,04619565	0,01358696	0	0,08967391	0	0,01902174	0,09782609	0	0,10597826
18	SP101558	Ovary-AdenoC	0,09205021	0,05020921	0,12552301	0	0	0	0,15062762	0,0209205	0,0041841	0,0167364	0,20502092	0,0209205	0	0,31380753
19	SP101564	Ovary-AdenoC	0,34453782	0,05042017	0,00840336	0,0210084	0,18487395	0,26470588	0	0,02521008	0,07563025	0	0,01260504	0,01260504	0	0
20	SP101572	Ovary-AdenoC	0,0964467	0	0	0,14720812	0,49746193	0,12690355	0,01015228	0	0,12182741	0	0	0	0	0
21	SP101576	Ovary-AdenoC	0,50968703	0,05514158	0,00298063	0,0119225	0,29657228	0,07302534	0,00298063	0,00447094	0,00596125	0	0,00447094	0,00745156	0,02533532	0
22	SP101580	Ovary-AdenoC	0,36895674	0,11704835	0,00508906	0,09414758	0,09669211	0,19847328	0,01017812	0	0,043257	0	0	0,03562341	0,01272265	0,0178117
23	SP101584	Ovary-AdenoC	0,13565891	0,15116279	0,01937984	0,1124031	0,1744186	0	0,07364341	0,04263566	0,16666667	0,03488372	0,00387597	0,08139535	0,00387597	0
24	SP101588	Ovary-AdenoC	0,07112971	0,30125523	0,0251046	0,0041841	0	0,07949791	0,0460251	0,08786611	0,13389121	0,07112971	0,0292887	0,12133891	0,0292887	0
25	SP101592	Ovary-AdenoC	0,29076087	0,00543478	0	0,08695652	0,27717391	0,27717391	0,00271739	0	0,05978261	0	0	0	0	0
26	SP101596	Ovary-AdenoC	0,23931624	0,21367521	0,02136752	0,0982906	0,18376068	0	0,01282051	0	0,08974359	0,00854701	0,01709402	0,03418803	0,05982906	0,02136752
27	SP101600	Ovary-AdenoC	0	0,07692308	0,23776224	0	0,02797203	0,01398601	0,30769231	0,0979021	0	0	0,17482517	0,04195804	0	0,02097902
28	SP101604	Ovary-AdenoC	0	0,05063291	0,03797468	0	0,03797468	0,03164557	0,05063291	0,05063291	0,03797468	0,22151899	0,25949367	0,06329114	0	0,15822785
29	SP101610	Ovary-AdenoC	0	0,05025126	0,06030151	0	0	0,06532663	0,04522613	0,04020101	0	0,0201005	0,15075537	0,0201005	0,09547739	0,27135678
30	SP101616	Ovary-AdenoC	0,2737069	0,22844828	0,03448276	0,00215517	0,00215517	0,22413793	0,09267241	0,05387931	0,01939655	0	0,03017241	0,02155172	0,01724138	0
31	SP101622	Ovary-AdenoC	0,24570025	0,16953317	0	0,002457	0,20884521	0,16953317	0	0	0,1007371	0	0,02211302	0,02457002	0	0,05651106
32	SP101628	Ovary-AdenoC	0,24099723	0,16897507	0	0	0,30193906	0,10526316	0,01108033	0	0,08310249	0,01108033	0	0,05540166	0	0,02216066
33	SP101634	Ovary-AdenoC	0	0,01748252	0,04895105	0	0	0	0,07342657	0,38811189	0,02097902	0,01048951	0,23776224	0,04195804	0	0,16083916
34	SP101642	Ovary-AdenoC	0,1875	0,1875	0,01785714	0	0,01636905	0,33184524	0,04464286	0,02232143	0,04166667	0	0,03720238	0,04017857	0	0,07291667
35	SP101648	Ovary-AdenoC	0	0,06565657	0,09090909	0	0,00505051	0,11111111	0,18181818	0,04040404	0	0,15656566	0,20707071	0,08080808	0	0,06060606
36	SP101654	Ovary-AdenoC	0,55952381	0,16369048	0,04166667	0	0,01488095	0,12797619	0,06547619	0	0,02083333	0	0,00595238	0	0	0
37	SP101658	Ovary-AdenoC	0	0,17460317	0	0	0	0	0,0952381	0,20634921	0,03174603	0	0,23809524	0,14285714	0,03174603	0,07936508

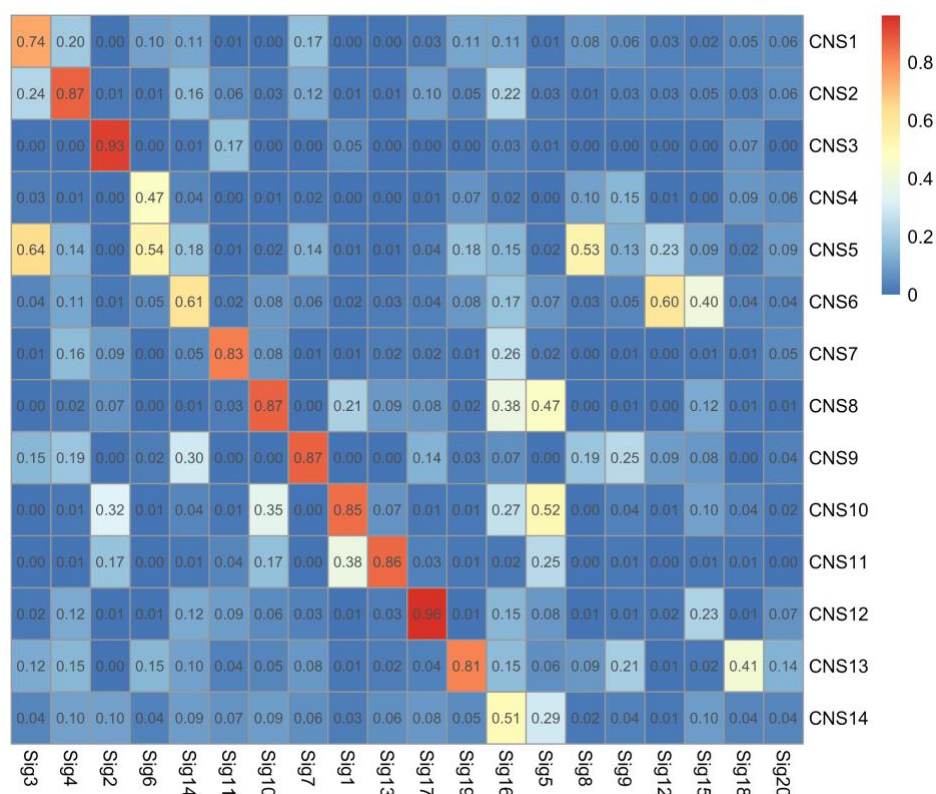
### 3.3. Confronto signatures PCAWG e TCGA

Per valutare la robustezza delle *signatures* estratte, sono state comparate tra loro quelle derivate dal database PCAWG e da TCGA.

Tale comparazione è stata eseguita in base alle similarità tra le *signatures*, ottenute utilizzando le medie della *cosine distance* (R) tra le *signatures*, a loro volta calcolate tramite la funzione *get\_sig\_similarity* presente nel pacchetto *Sigminer*.

A partire dal grafico ottenuto (*Figura 5*), è possibile osservare che:

1. La maggior parte delle *signatures* PCAWG (9/14) sono correlate alle *signatures* TCGA, in quanto presentano  $R \geq 0.8$ .
2. Quattro *signatures* PCAWG (CNS4, CNS5, CNS6 e CNS14) hanno similarità intermedia ( $R \geq 0.51$ ) con la controparte TCGA. Ciò potrebbe avere due cause: non sono *signatures* condivise tra vari tipi tumorali oppure hanno dei pattern differenti tra i due dataset.
3. La maggior parte delle *signatures* TCGA (16/20) hanno R medio alto rispetto alle *signatures* PCAWG.
4. Le quattro *signatures* TCGA rimanenti (Sig15, Sig18, Sig9 e Sig20) non sono correlate con nessuna *signature* PCAWG. Probabilmente queste *signatures* sono rappresentate solo in un particolare tessuto.



**Figura 5:** correlazione tra le signatures estratte da PCAWG (nelle ascisse) e TCGA (nelle ordinate). Valori tendenti al rosso indicano una maggiore correlazione, mentre ai valori tendenti al blu indicano una bassa correlazione. Vengono inoltre indicati i valori di cosine similarity per ogni comparazione.

È possibile notare, inoltre, che le due tipologie di *signatures* hanno un andamento “a coppie”. Ciò indica che la funzione utilizzata per l'estrazione (*SigProfilerExtractor*) e le funzioni presenti in *Sigminer* sono degli ottimi metodi di analisi delle *signatures*.

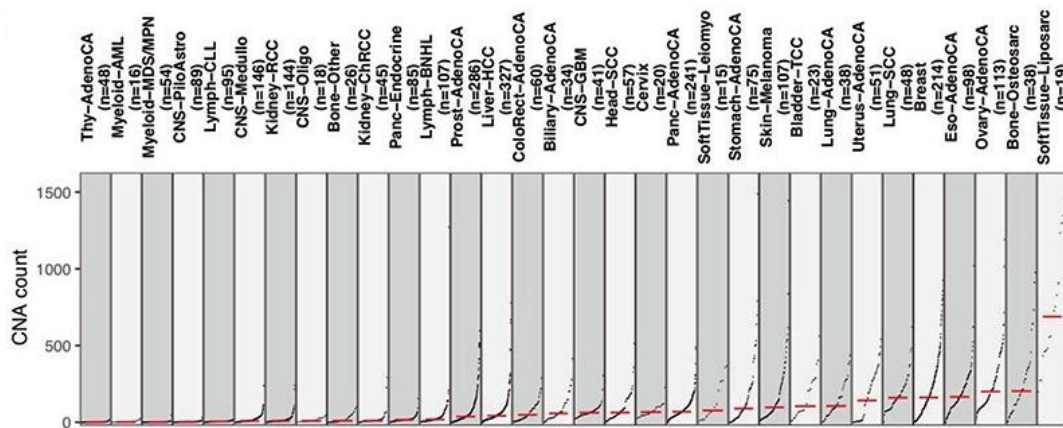
Le incongruenze presenti in questo andamento sono causate principalmente da:

- La tecnica di biologia molecolare utilizzata per la generazione dei valori di CN: a partire da SNP per TCGA e a partire da WGS per PCAWG.
- L'uso di *software* diversi per analizzare le CNV. Dato che i valori di CN ottenuti da PCAWG sono ricavati da molti *software* diversi, ci si aspetta che essi abbiano una qualità molto più alta rispetto ai valori di TCGA.

### 3.4. Distribuzione delle *signatures pan-cancer* e correlazione con i processi mutagenici

Nel dataset PCAWG, il numero medio di segmenti con CNA per ogni paziente è circa due.

Alcuni tipi di cancro hanno un numero di segmenti minore, come la leucemia mieloide acuta (AML), mentre altri ne hanno un numero maggiore, come il cancro all'ovaio (Ovary-AdenoCA) e al seno (Breast) (*Figura 6*).



*Figura 6: numero di CNA per tipo di cancro*

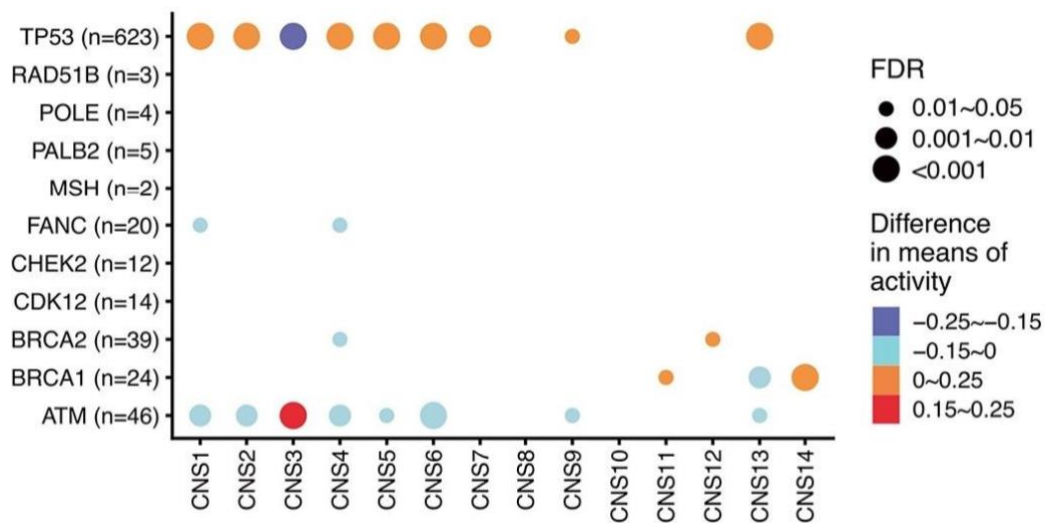
Un importante scopo dell'analisi delle *signatures* è l'identificazione dei processi mutagenici che portano all'alterazione del CN.

Sono state quindi categorizzate, basandosi sul dataset PCAWG, le varianti genetiche presenti nelle cellule germinative (spermatozoi e ovuli) che presentano il rischio di sviluppare il cancro, e le mutazioni *driver*, ovvero su geni coinvolti nella riparazione del DNA.

Successivamente si sono cercate delle correlazioni con l'attività delle *signatures*.

Sono stati ottenuti i seguenti risultati:

1. Le mutazioni funzionali dell'oncogene *BRC1A1* sono attribuibili a CNS14 (*Figura 7*). Dato che CNS14 presenta una forte correlazione con le *signatures* di COSMIC SBS3 e ID6, le quali sono note per la loro associazione con l'incapacità di riparare il DNA tramite il meccanismo di ricombinazione omologa (HRD Homologous Recombination Deficiency), ciò suggerisce che CNS14 sia associato con tale meccanismo mutagenico.



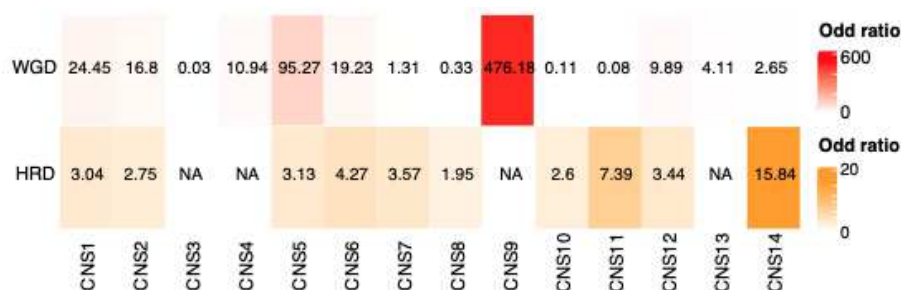
**Figura 7:** associazione delle mutazioni di geni chiave nella riparazione del DNA con le attività delle signatures.

Per ogni gene i campioni derivanti dal dataset PCAWG sono stati divisi in due gruppi in base allo stato di mutazione del gene. I valori delle attività relative nei due gruppi sono stati differenziati tramite l'utilizzo di colori diversi (rosso e viola).

La grandezza dei punti indicano i P-values corretti in base alla presenza di falsi positivi (FDR - False Discovery Rate).

Il numero di fianco al gene si riferisce al numero di mutazioni patogeniche, sia geminali che somatiche, riscontrabili nel gene in questione.

2. L'attività della *signature* CNS9 fornisce una previsione molto accurata del WGD, suggerendo che potrebbe essere un indicatore affidabile dello stato di duplicazione (Figura 8).
3. L'attività della *signature* CNS14 fornisce una previsione molto accurata di HRD (Figura 8).



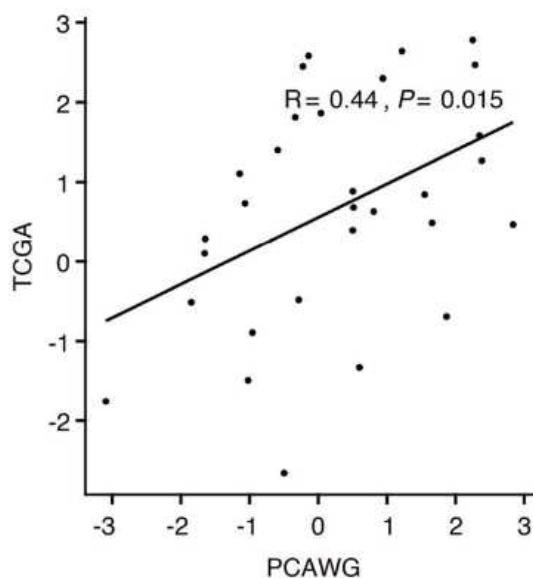
**Figura 8:** associazione tra WGD, stato di HRD e signatures.

### 3.5. Analisi cliniche

Per testare l'ipotesi che le *signatures* estratte possano essere utilizzate come *biomarker* per la prognosi della malattia, è stata eseguita un'analisi applicando il modello a rischi proporzionali di *Cox* per ogni tipo di cancro.

Gli *Hazard Risk* calcolati (*Figura 9*) indicano che le attività delle *signatures* estratte tra i due database sono strettamente correlate ( $R = 0.44$  con la correlazione di Pearson).

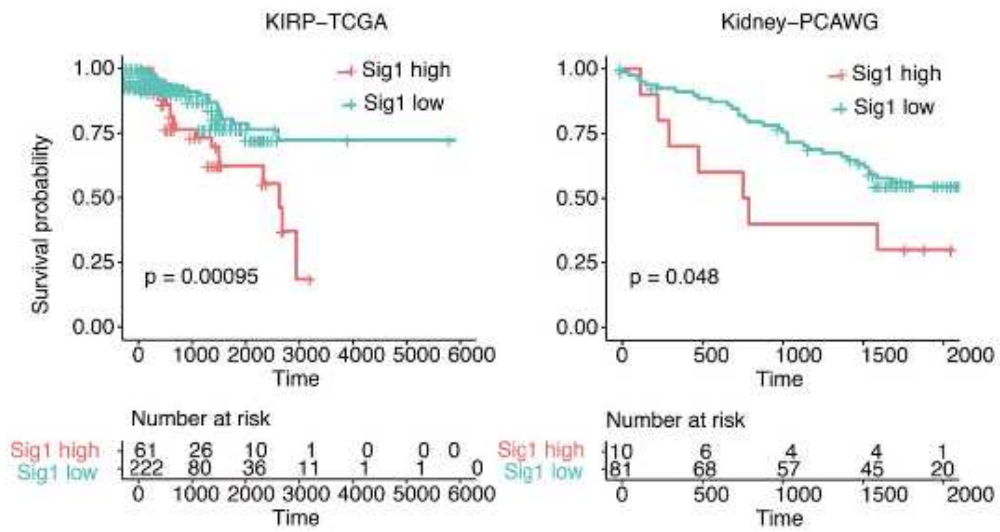
Tale risultato dimostra che l'attività della *signatures* può essere un valido indicatore nel prevedere della sopravvivenza del paziente.



**Figura 9:** Correlazione tra gli Hazard Risk dei tipi di cancro tra i due dataset. Viene riportato inoltre il coefficiente di Pearson

Analizzando il grafico di *Kaplan-Meier* in *figura 11*, è possibile osservare che, per esempio, nel cancro al rene un'alta attività di *signature 1* è associata con una significativa riduzione della sopravvivenza, sia nel dataset TCGA che in PCAWG.





**Figura 10:** grafici di Kaplan-Meier che mostrano la comparazione della sopravvivenza tra i pazienti che presentano un'attività maggiore del valore soglia (high, curva rossa) e un'attività minore (low, curva verde). Il valore soglia è stato calcolato tramite la funzione `surv_cutpoint` presente nel pacchetto `survminer`.

In basso è inoltre presente la tabella di rischio, indicante il numero di pazienti ancora coinvolti nello studio ad un momento temporale specifico (in giorni).

#### 4. DISCUSSIONE

Riassumendo i risultati ottenuti, è possibile stabilire che il metodo per l'analisi delle *signatures* sviluppato da Tao *et al.* può essere applicato per pazienti oncologici che presentano *signatures* generate a partire da WGS o SNP con buoni risultati.

Sono stati infatti identificati nuovi *pattern* di CNA ed è stato dimostrato che è presente una correlazione significativa tra l'attività di alcune *signatures* e la prognosi dei pazienti analizzati.

Un vantaggio di questo metodo consiste nella rilevazione di nuovi *pattern* di CNA, cosa non possibile con i metodi di analisi precedentemente esistenti in quanto si basano sull'identificazione delle *signatures* a partire da quelle presenti nei database quali COSMIC.

Tramite il metodo sviluppato da Tao *et al.*, è infatti possibile osservare come i nuovi *pattern* estratti non derivino da un unico evento mutagenico, bensì da una combinazione di diversi eventi. Per esempio, CNS6 potrebbe essere una combinazione di un aneuploidia (corrispondente a CNS11) e WGD (CNS9).

Un ulteriore vantaggio consiste nella vasta possibilità di applicazione, infatti in questo studio sono state eseguite analisi su dati estratti da SNP e WGS.

In aggiunta, la comparazione dei profili di *signatures* presenti in due dataset (TCGA e PCAWG) ha permesso un'ulteriore validazione dei risultati ottenuti, in particolare riguardo le analisi di sopravvivenza.

Tuttavia, l'esplorazione delle *signatures* di CNV è ancora nelle fasi iniziali. Dal lavoro di Tao *et al.* emergono infatti delle questioni che necessitano di ulteriori analisi.

Nonostante uno degli obiettivi principali dell'analisi delle *signatures* di CNV sia rivelare i processi mutagenici alla base delle *signatures* stesse, tali processi rimangono per la maggior parte sconosciuti e richiedono ulteriori studi e approfondimenti.

Concludendo, questo lavoro getta dunque le basi per studi futuri riguardanti l'eziologia delle *signatures* e la realizzazione di *biomarkers* affidabili per la diagnosi precisa della patologia e per l'elaborazione di un appropriato piano terapeutico.

## 5. BIBLIOGRAFIA

- [1] X. Shao *et al.*, «Copy number variation is highly correlated with differential gene expression: a pan-cancer study», *BMC Medical Genetics*, vol. 20, fasc. 1, p. 175, nov. 2019, doi: 10.1186/s12881-019-0909-5.
- [2] C. D. Steele, N. Pillay, e L. B. Alexandrov, «An overview of mutational and copy number signatures in human cancer», *J Pathol*, vol. 257, fasc. 4, pp. 454–465, lug. 2022, doi: 10.1002/path.5912.
- [3] O. Pich, F. Muiños, M. Paul Lolkema, N. Steeghs, A. Gonzalez-Perez, e N. Lopez-Bigas, «The mutational footprints of cancer therapies», *Nat Genet*, vol. 51, fasc. 12, pp. 1732–1740, dic. 2019, doi: 10.1038/s41588-019-0525-5.
- [4] S. M. A. Islam *et al.*, «Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor», *Cell Genomics*, vol. 2, fasc. 11, p. 100179, nov. 2022, doi: 10.1016/j.xgen.2022.100179.

## 6. APPENDICE 1: *Esperienza di Stage*

### 6.1. Studi analizzati

Al termine del mio percorso triennale ho avuto la possibilità di eseguire un tirocinio di 150 ore presso il laboratorio della Professoressa Romualdi.

Nel corso dello stage ho confrontato tre studi riguardanti la tematica delle *signatures* di CNV:

1. **Z. Tao *et al.***, «The repertoire of copy number alteration signatures in human cancer», *Briefings in Bioinformatics*, vol. 24, fasc. 2, p. bbad053, mar. 2023, descritto in questo elaborato
2. **C. D. Steele *et al.***, «Signatures of copy number alterations in human cancer», *Nature*, vol. 606, fasc. 7916, Art. fasc. 7916, giu. 2022
3. **M. Drews *et al.***, «A pan-cancer compendium of chromosomal instability», *Nature*, vol. 606, fasc. 7916, Art. fasc. 7916, giu. 2022

Tutti gli studi partono da set di campioni diversi, ma con molti campioni in comune, ed estraggono le *signatures* a partire da essi tramite la funzione *SigProfilerExtractor*.

### 6.2. Metodi utilizzati

Inizialmente sono partito dallo studio di Steel *et al.* e ho eseguito i passaggi analoghi a quelli descritti nei [Metodi](#) in questo elaborato per estrarre le *signature*. In particolare, a partire dal dataset di circa 11'000 campioni TCGA, ho generato in Python la matrice di campioni tramite il tool *SigProfilerMatrixGenerator* e, a partire da quest'ultima, ho estratto le *signatures* tramite *SigProfilerMatrixGenerator*.

Sono state riscontrate delle difficoltà in quest'ultimo passaggio in quanto la documentazione fornita da Steel *et al.* non era sufficientemente dettagliata e la funzione forniva numerosi errori.

Dopo numerosi tentativi, sono state estratte 16 *signatures*, nominate da CNV48A a CNV48Z.

Tale passaggio permette di confrontare tra loro le *signatures* estratte dai tre studi sopra citati e quelle estratte *de novo*.

A partire da tali *signatures* ho ricavato i campioni in comune (5382 pazienti) ed eseguito alcune analisi.

Inizialmente ho calcolato la correlazione tra le *signatures* dei vari lavori tramite la funzione *cor* presente nella libreria R *stats*. Le correlazioni sono state

successivamente rappresentate in un grafico tramite la funzione *corrplot*, presente nel pacchetto R *corrplot*.

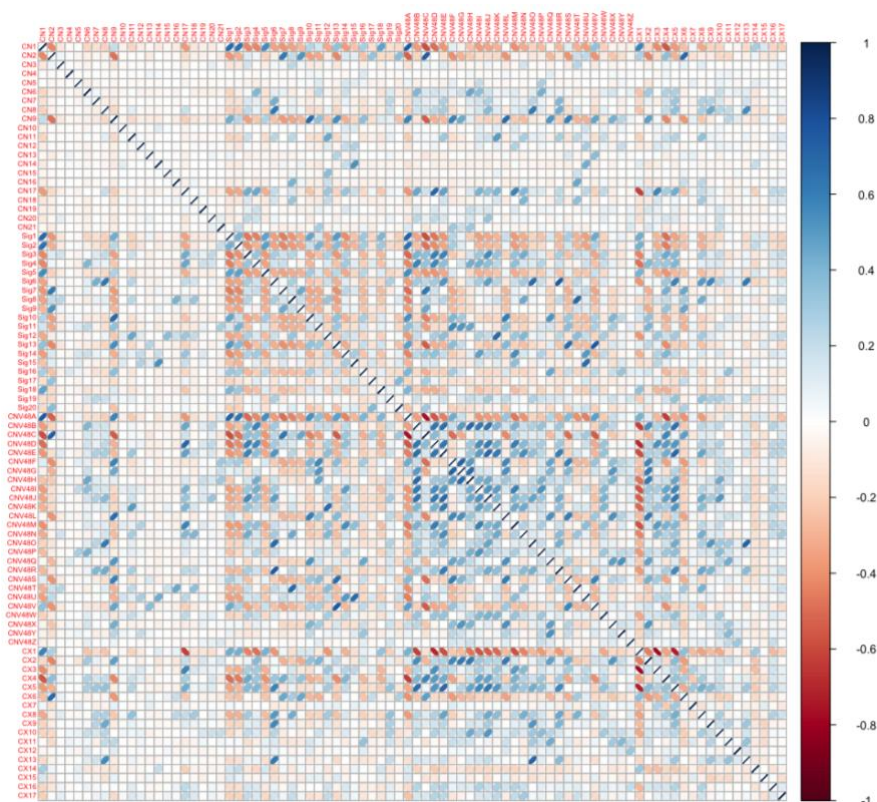
In seguito ho realizzato una *heatmap* tra tutte le *signatures* e i campioni, clusterizzando quest'ultimi. Ho inoltre aggiunto una colonna rappresentante i tipi tumorali per ogni paziente. Tale *heatmap* è stata realizzata tramite la funzione *Heatmap* presente nel pacchetto R *ComplexHeatmap*.

Infine ho eseguito le analisi cliniche sulle *signatures* cercando di ricavare quali di esse abbiano un effetto sulla sopravvivenza dei pazienti. Tali analisi sono state eseguite tramite la funzione *coxph* (presente nel pacchetto *survival* in R) e, a partire da quest'ultime, è stato realizzato un grafico di *Kaplan-Meier* tramite la funzione *surv\_fit* contenuta nel pacchetto R *survminer*.

### 6.3. Risultati ottenuti

#### 6.3.1. Correlazioni

Dalla *Figura 11* è possibile osservare che le *signatures* estratte da Steel *et al.* hanno un ottimo andamento in quanto sono poco correlate tra loro. Inoltre, le *signatures* CN1, CN9 e CN17 sembrano essere correlate con molte *signatures* degli altri studi.



**Figura 11:** grafico delle correlazioni tra le *signatures*. Il colore delle ellissi rappresenta la forza della correlazione mentre la direzione indica se si tratta di una correlazione negativa o positiva. Da sinistra verso destra e dall'alto verso il basso sono rappresentate rispettivamente le *signatures* di: Steel, Tao, estratte de novo e Drews



### 6.3.3. Analisi Cliniche

Le analisi cliniche sono state eseguite su due variabili: sopravvivenza (OS - Overall Survival) e *Progression-Free Interval* (PFI), ovvero il tempo dall'assegnazione ad un *trial* clinico o dall'inizio della terapia alla progressione della malattia o alla morte.

Dopo aver utilizzato la funzione *coxph* sono state ricavate 35 *signatures* che impattano la sopravvivenza del paziente.

Le *signatures* che, quando attive, portano ad aumento della probabilità di sopravvivenza sono:

- CN2, CN7 (*Figura 14* in Appendice 2), CN14, di Steel *et al.*
- CX14 di Drews *et al.*

Le *signatures* che, invece, una volta attive, portano ad una diminuzione delle probabilità di sopravvivenza sono:

- CN8, CN9, CN10 di Steel *et al.*
- Sig4, Sig10, Sig13, Sig16, di Tao *et al.*
- CNV48B, CNV48K, CNV48O, CNV48W di quelle estratte *de novo*
- CX2 (ma non a lungo termine), CX3, CX5, CX13 (solamente a breve termine) di Drews *et al.*

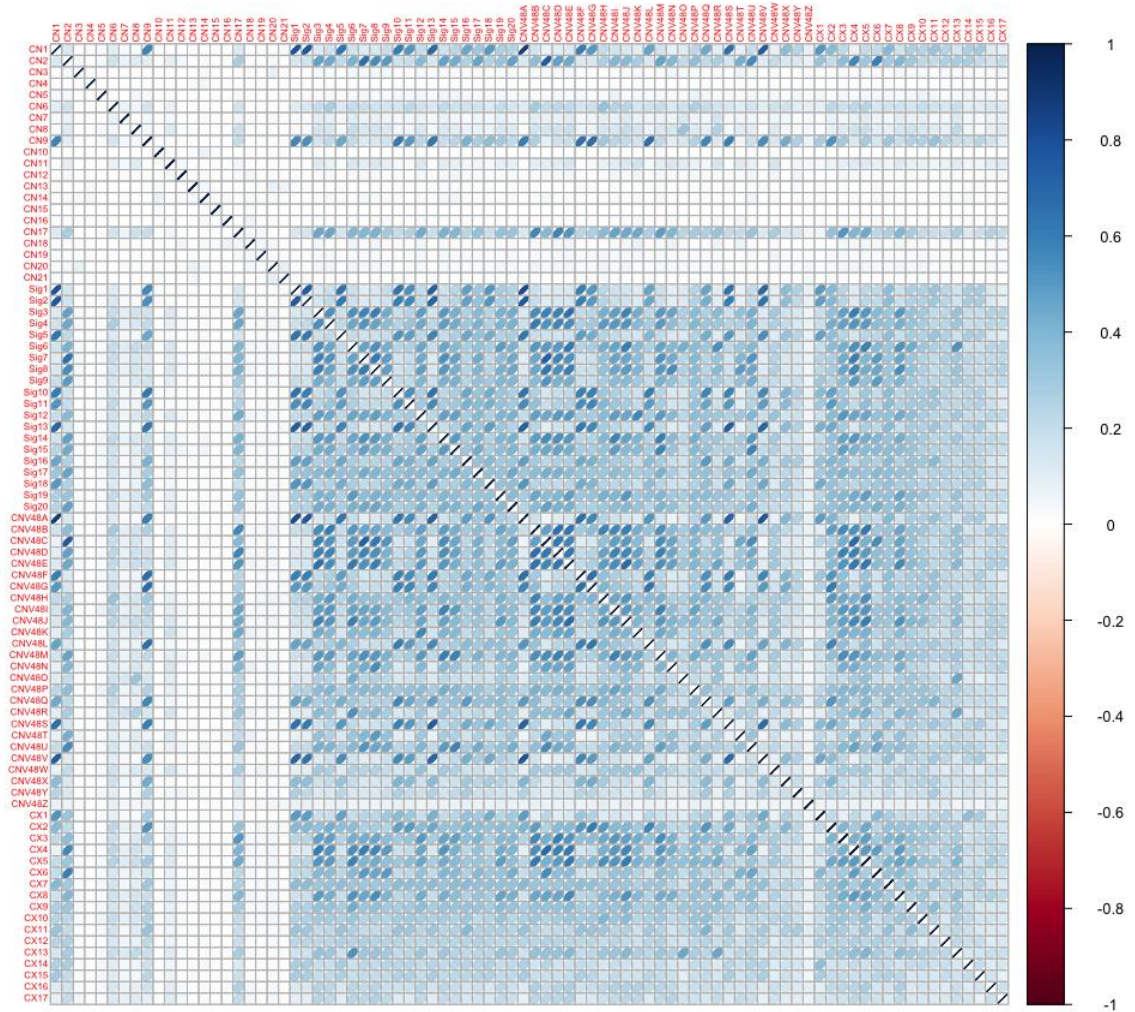
Le *signatures* che, una volta attive, portano ad aumento del tempo di PFI sono:

- CN2, CN7 (*figura 15* in Appendice 2), CN11, CN14,
- CX1,
- CNV48Z

Le *signatures* che, invece, una volta attive, portano ad una diminuzione del tempo di PFI sono:

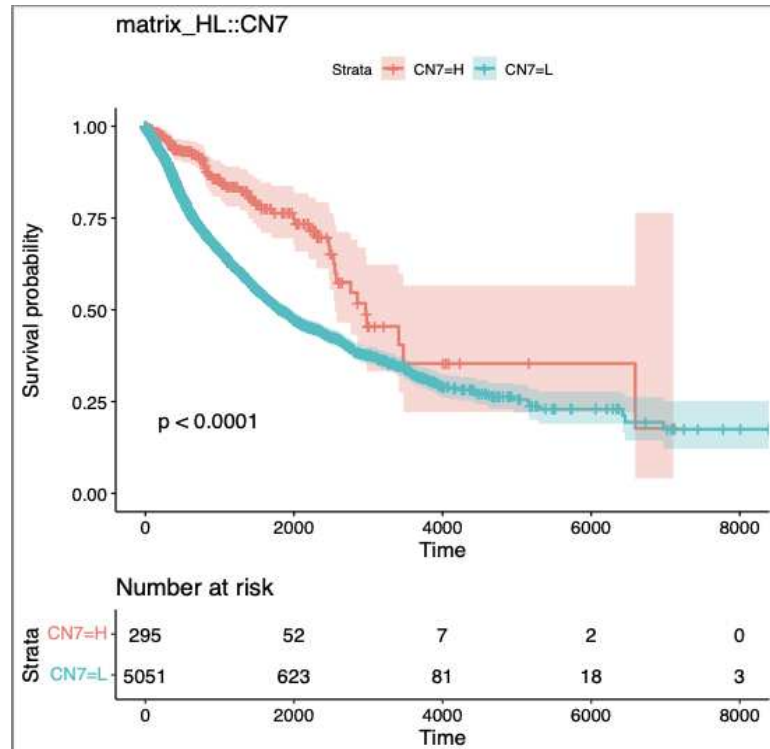
- CN8, CN9, CN10, CN17 di Steel *et al.*
- Sig4, Sig13, Sig14, Sig15, Sig16 di Tao *et al.*
- CX3, CX5, CX13 di Drews *et al.*
- CNV48A, CNV48B, CNV48I, CNV48K, CNV48Q, CNV48X di quelle estratte *de novo*

## 7. APPENDICE 2: *Grafici Supplementari*



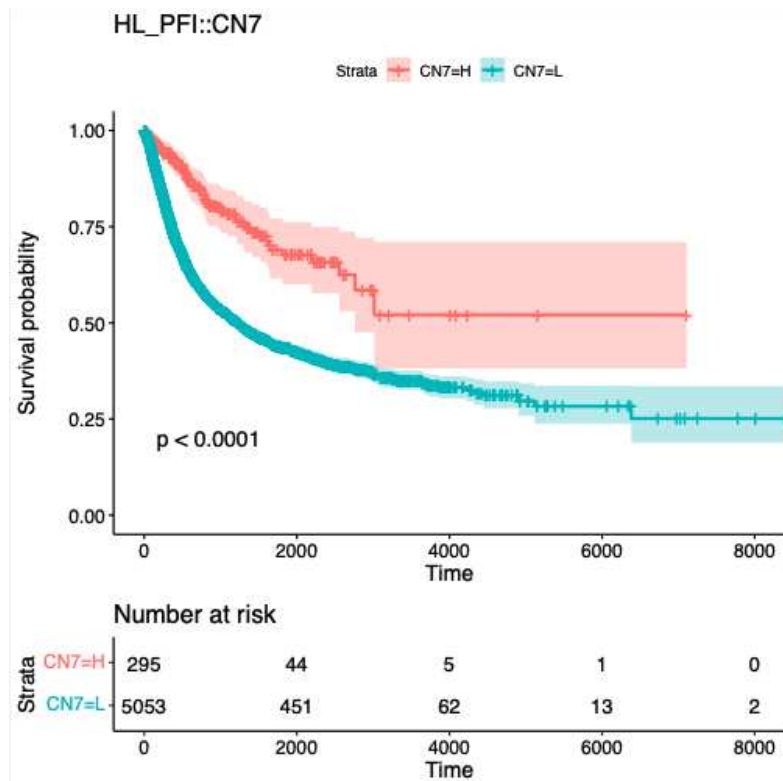
**Figura 13:** grafico delle correlazioni tra le signatures calcolato tramite l'indice di Jaccard con la funzione "Jaccard" presente nel pacchetto R "Signac". È possibile osservare un andamento analogo alla figura 11.





**Figura 14:** grafico di Kaplan-Meier per la signature CN7 di Steel et al. sull'analisi del tempo di sopravvivenza.





La linea azzurra (L - low) indica una bassa attività della signature mentre la linea rossa (H - high) rappresenta un'alta attività della signature.



**Figura 15:** grafico di Kaplan-Meier per la signature CN7 di Steel et al. sull'analisi del tempo di PFI.

La linea azzurra (L - low) indica una bassa attività della signature mentre la linea rossa (H - high) rappresenta un'alta attività della signature.

# The repertoire of copy number alteration signatures in human cancer

Ziyu Tao <sup>†</sup>, Shixiang Wang <sup>†</sup>, Chenxu Wu<sup>†</sup>, Tao Wu , Xiangyu Zhao, Wei Ning, Guangshuai Wang, Jinyu Wang, Jing Chen, Kaixuan Diao, Fuxiang Chen and Xue-Song Liu 

Corresponding author: Xue-Song Liu. School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. Tel: +86-21-20684520; Fax: +86-21-20685430; E-mail: [liuxs@shanghaitech.edu.cn](mailto:liuxs@shanghaitech.edu.cn)

<sup>†</sup>Ziyu Tao, Shixiang Wang and Chenxu Wu contributed equally to this work.

## Abstract

Copy number alterations (CNAs) are a predominant source of genetic alterations in human cancer and play an important role in cancer progression. However comprehensive understanding of the mutational processes and signatures of CNA is still lacking. Here we developed a mechanism-agnostic method to categorize CNA based on various fragment properties, which reflect the consequences of mutagenic processes and can be extracted from different types of data, including whole genome sequencing (WGS) and single nucleotide polymorphism (SNP) array. The 14 signatures of CNA have been extracted from 2778 pan-cancer analysis of whole genomes WGS samples, and further validated with 10 851 the cancer genome atlas SNP array dataset. Novel patterns of CNA have been revealed through this study. The activities of some CNA signatures consistently predict cancer patients' prognosis. This study provides a repertoire for understanding the signatures of CNA in cancer, with potential implications for cancer prognosis, evolution and etiology.

**Keywords:** copy number alteration, mutational signature, cancer genome, cancer prognosis, copy number signature

## Background

Somatic mutations are the driving force of cancer development. Genomic alterations in cancer cells consist of two major categories: (i) small scale alterations that include single base substitutions (SBSs) and small insertion and deletions (INDELs) and (ii) large scale alterations known as structural variations (SV). Copy number alteration (CNA) is a major type of SV and is prevalent in human cancer [1, 2]. CNAs and SBSs stem from distinct mutational processes, SBSs are usually caused by lesions or repair mistakes in single-strand deoxyribonucleic acid (DNA), while CNAs are the results of double-strand DNA breaks, double-strand DNA-repair defects, DNA replication or cell division defects [3].

CNAs have critical roles in activating oncogenes and in inactivating tumor suppressors [4, 5]. Additionally, aneuploidy status can influence cancer cell proliferation and competitiveness [6, 7]. In morphologically normal tissues, similar SBS have been observed as in cancer cells; however, CNAs are mostly observed

in cancer cells but not in morphologically normal tissues [8]. CNAs have been reported to predict cancer relapse and prognosis [9, 10]. These observations suggest that CNAs play a critical role in the malignant transformation of normal cells to cancer cells. However, the underlying mechanism is largely unknown.

Genomic DNA alteration signatures are recurring genomic patterns that are the imprints of mutagenic processes accumulated over the lifetime of cancer cell [11, 12]. Genome alteration signature analysis can not only provide the mutational process information but also biomarkers for cancer precision medicine [13, 14]. SBS signature analysis has been extensively studied, and represents a prototype for other types of signature study [12]. Despite the importance of CNA in cancer progression, a comprehensive understanding of the mutational process and signature of CNA is still lacking.

Signatures of SV have been studied in breast cancer [15]. However, this method relies on high coverage Whole Genome Sequencing (WGS) data and cannot be applied with SNP array or whole

Ziyu Tao is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [taozy@shanghaitech.edu.cn](mailto:taozy@shanghaitech.edu.cn)

Shixiang Wang is a research fellow at Department of Experimental Research, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Sun Yat-sen University, Guangzhou 510060, China. [wangshx@shanghaitech.edu.cn](mailto:wangshx@shanghaitech.edu.cn)

Chenxu Wu is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [wuchx@shanghaitech.edu.cn](mailto:wuchx@shanghaitech.edu.cn)

Tao Wu is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [wutao2@shanghaitech.edu.cn](mailto:wutao2@shanghaitech.edu.cn)

Xiangyu Zhao is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [ZHAOxy2@shanghaitech.edu.cn](mailto:ZHAOxy2@shanghaitech.edu.cn)

Wei Ning is a master student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [ningwei@shanghaitech.edu.cn](mailto:ningwei@shanghaitech.edu.cn)

Guangshuai Wang is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [wanggsh@shanghaitech.edu.cn](mailto:wanggsh@shanghaitech.edu.cn)

Jinyu Wang is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [wangjy10@shanghaitech.edu.cn](mailto:wangjy10@shanghaitech.edu.cn)

Jing Chen is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [chenjing2@shanghaitech.edu.cn](mailto:chenjing2@shanghaitech.edu.cn)

Kaixuan Diao is a PhD student at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [diaokx@shanghaitech.edu.cn](mailto:diaokx@shanghaitech.edu.cn)

Fuxiang Chen is a professor at Department of Clinical Immunology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 200011, People's Republic of China. [chenfx@sjtu.edu.cn](mailto:chenfx@sjtu.edu.cn)

Xue-Song Liu is a professor at School of Life Science and Technology, ShanghaiTech University, Shanghai 201203, China. [liuxs@shanghaitech.edu.cn](mailto:liuxs@shanghaitech.edu.cn)

Received: June 15, 2022. Revised: January 1, 2023. Accepted: January 26, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

exome sequencing (WES) data. Macintyre *et al.* employed a mixture modeling based method for copy number (CN) component extraction [16]. For each different dataset or cancer type, a different set of CNA components will be generated based on the distributions of the six CNA features (segment size, change point CN and segment CN, breakpoint count per 10 Mb, length of segments with oscillating CN and breakpoint count per chromosome arm). This inconsistency of the CNA components among different datasets limits the generalization of their method in different cancer types or different datasets. Steele *et al.* reported a method for CNA signature analysis with 48 features, which considered the absolute CN, size and heterozygosity status of CNA segment; however, the background information of the CNA segment has not been incorporated [17]. We recently developed a pre-defined set of CNA components and corresponding software implementation as a module of Sigminer (<https://cran.r-project.org/package=sigminer>) for CNA signature analysis [18–20]. This set of CNA components incorporates the six reported CNA features and includes two additional features. The application of this tool (Sigminer) in prostate cancer reveals distinct CNA mutational processes and clinical outcomes [18]. The Macintyre *et al.* method, Steele *et al.* method and our recent method have a limited number of CNA features, a unified and comprehensive CNA classification method across different cancer types is still lacking.

Here we developed a mechanism-agnostic method for CN segment categorization and signature extraction. Our method incorporates the following information for each DNA segment: absolute CN, CN context, segment length and loss of heterozygosity (LOH) status. The selection of these CN features was inspired by known patterns of CNA, such as chromothripsis and whole genome duplication (WGD) [3, 21, 22]. With this new CN signature analysis method, a pan-cancer landscape of CNA signature is shown. Known CNA patterns have been reproduced, and new CNA signatures have been identified in this pan-cancer study, such as haploid chromosome and combination with WGD. Underlying mutational processes for the identified CNA signatures have been investigated. The activities of some CNA signatures consistently predict cancer patients' prognosis, suggesting CNA signatures could be cancer prognosis biomarkers.

## Results

### The pan-cancer landscape of CNAs

We used the WGS dataset from pan-cancer analysis of whole genomes (PCAWG) to study the profile of CNA. The CNA profiles are then validated with independent the cancer genome atlas (TCGA) SNP array dataset. PCAWG dataset contains the WGS (38–60X sequencing) data of 2778 samples (32 cancer types) [23]. TCGA dataset includes SNP array data (data platform: Affymetrix SNP 6.0) of 10 851 samples (33 cancer types) [1].

The length distribution of CNA in the PCAWG dataset and TCGA dataset are shown (Figure 1A). Similar to the previous observation [2], the focal CNAs occur at a frequency inversely related to their lengths, arm-level CNAs occur more frequently than would be expected by the inverse-length distribution associated with focal CNAs. This indicates that compared with focal CNAs, chromosome arm-level CNAs are generated through different mutational processes. CNA burden measures the percent of CN altered genome [9]. Pan-cancer distributions of CNA burden and CNA segment number are shown (Figure 1B and C, Supplementary Figure 1A and B).

Pan-cancer density distribution of the values of CNA burden and CNA count show multi-modal distribution (Figure 1D). In

PCAWG dataset, CNA burden and total CNA show a strong positive correlation, SBS and INDEL show a strong positive correlation, CNA and SBS, INDEL show a weak positive correlation (Supplementary Figure 1C). In TCGA dataset, similar correlations exist as in PCAWG dataset, except, CNA and INDEL show weak negative correlation (Supplementary Figure 1C). INDELS of TCGA dataset are derived from WES, while INDELS of PCAWG dataset are detected using WGS data, the difference in noncoding regions could contribute to this discordance in CNA and INDEL correlation. Pan-cancer distributions of these genome alteration features (CNA burden, Total CNA, Total INDEL, Total SBS) in PCAWG (Figure 1E) and TCGA (Supplementary Figure 1D) are shown. Some types of cancer show over-representation of CNA but not SBS, INDEL, such as breast, ovarian cancer, while cancer types including lung cancer show over-representation of all these types of genome alterations (Figure 1E).

### Design of CNA features and components

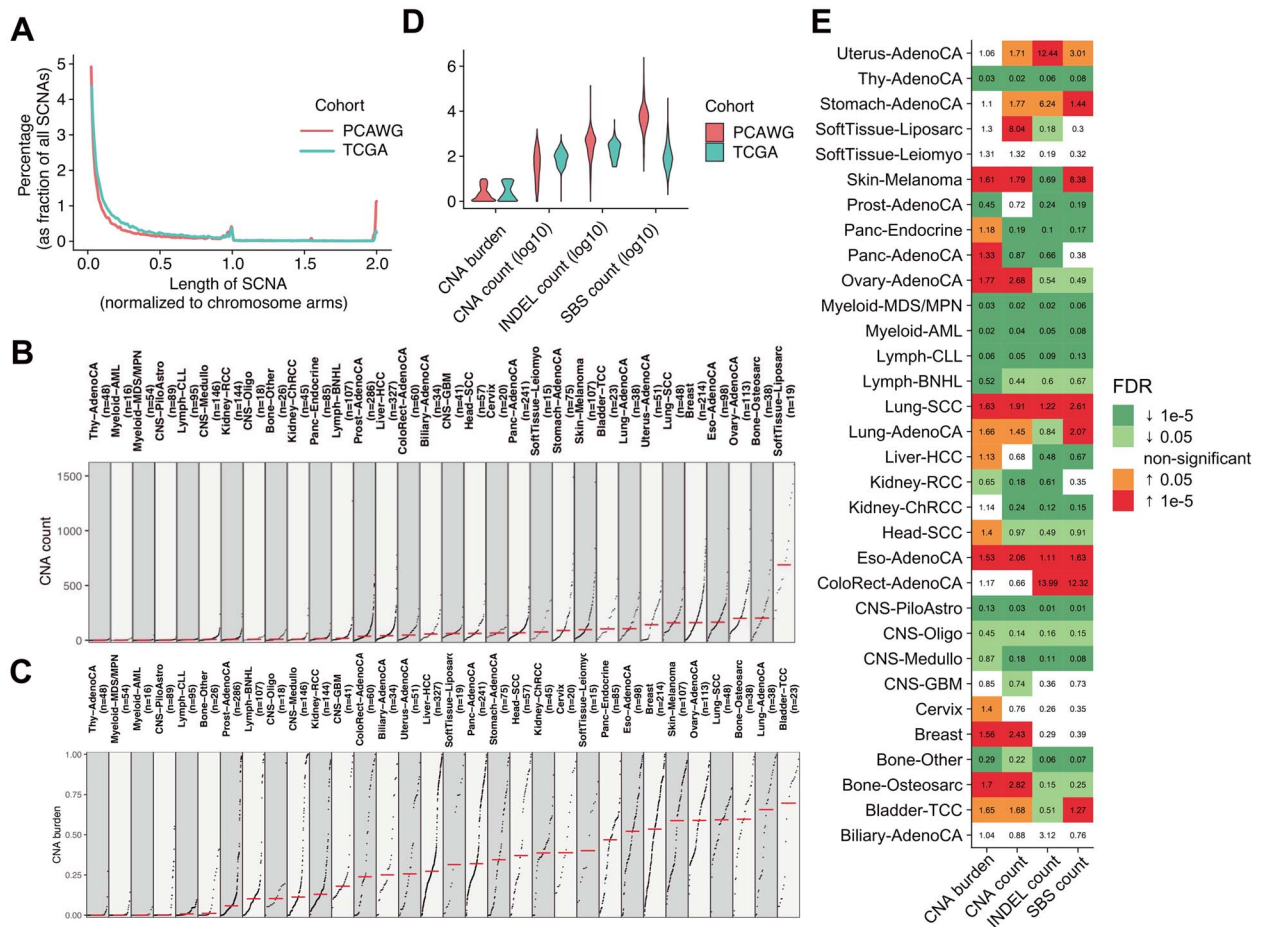
An essential step for CNA signature analysis is to design proper CNA features and components to classify CNAs. A clearly and stably defined set of CNA components is important for the generalization of CNA signature analysis in various types of cancer. Here we developed a new CNA classification method for CNA signature analysis. Our new classification method considers the following CNA features: morphology or context of CNA, absolute CN, LOH status. Furthermore, our method is applicable to different types of raw data, including WGS, WES, SNP array or panel sequencing data. Each CNA segment was classified by considering the following detailed features, (i) segment context, including segment shape composed of both the left and right segments of the target segment (low–low, high–high and ladder) and CN change number. In total, six segment context shapes have been defined (Figure 2). (ii) Absolute CN. Including the following components: 0, 1, 2, 3, 4, 5–8 and  $\geq 9$ . (iii) LOH status. (iv) Segment size, including the following components [24]: S (length < 50 kb); M (50 kb  $\leq$  length < 500 kb); L (500 kb  $\leq$  length < 5 Mb); E (5 Mb  $\leq$  length). In total, 176 components have been defined to characterize the CNA segments of human cancer patients (Supplementary Table 1).

### CN signature extraction in pan-cancer datasets

Based on the features and components of CNA segment defined above, for each cancer sample, the values for each CNA component will be calculated from the absolute CN profiles derived from WGS or SNP array data.

A CN component value matrix was generated by combining component values in all tumors. This matrix was subjected to non-negative matrix factorization (NMF), a method previously used for deriving SBS signatures [11]. For de novo CN signature extraction, we applied the widely used tool SigProfiler (Supplementary Figure 2) [12]. SigProfiler has also been used for the extraction of the standard SBS, doublet base substitutions (DBS) and small insertions and deletions (ID) mutational signatures stored in the catalog of somatic mutations in cancer (COSMIC) compendium [12].

As reported previously, the choice of the number of mutational signatures is rarely amenable to complete automation [12]. Here, the number of signatures extracted was determined using two parameters. First is the reconstruction error and the average Frobenius reconstruction error is reported. Second is the stability of signature extraction and the cosine similarity between the extracted signatures, average silhouette width, is reported (Supplementary Figure 3). In the 2778 PCAWG WGS dataset, 14 CNA signatures have been extracted, these signatures



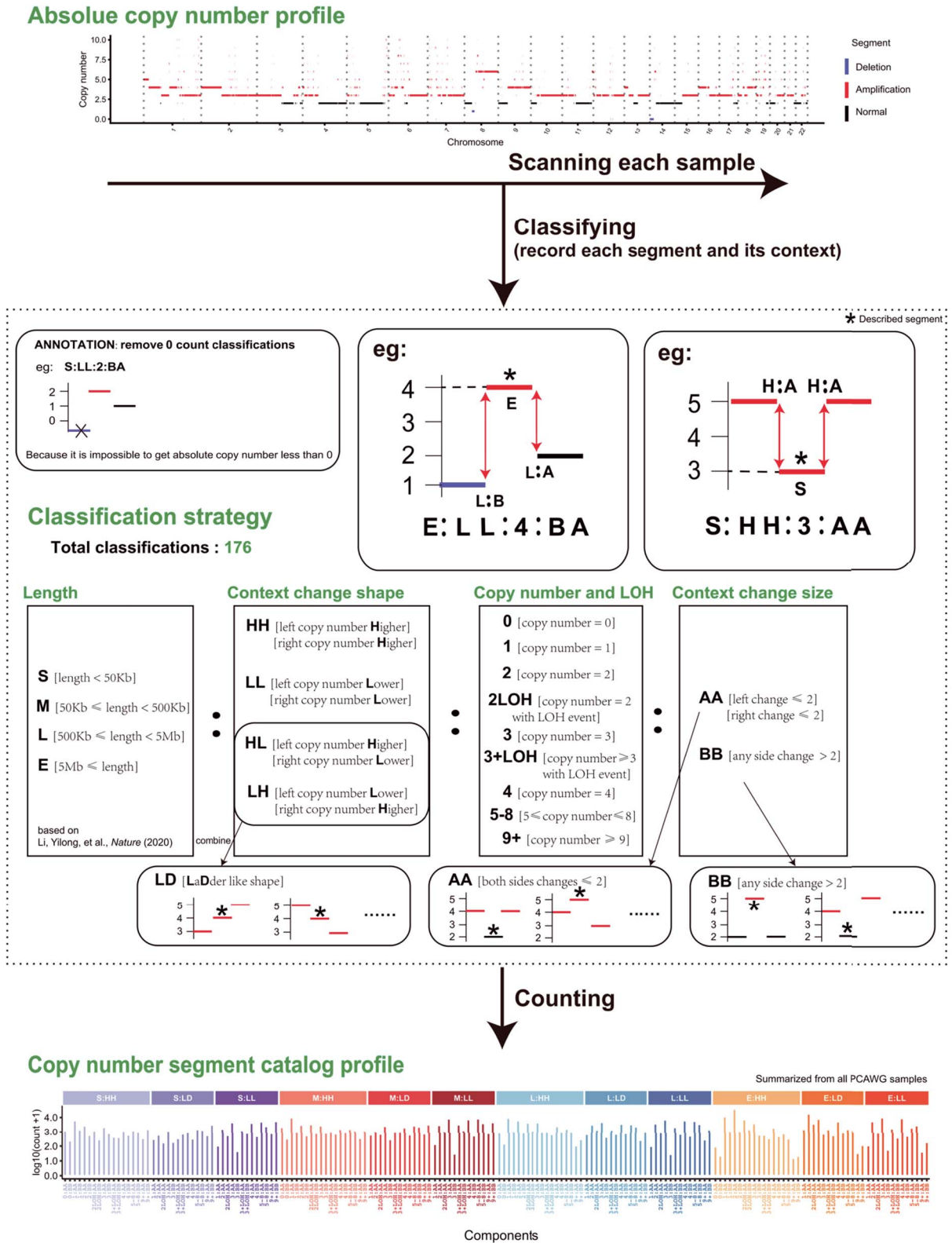
**Figure 1.** Pan-cancer distribution patterns of CNA. **(A)** Length distribution of CNA in PCAWG (red line) and TCGA (blue line) pan-cancer datasets. **(B, C)** Pan-cancer distribution pattern of CNA count **(B)**, CNA burden **(C)** in individual tumors of PCAWG dataset. **(D)** Pan-cancer distribution of the values of CNA burden, CNA count, INDEL count and SBS count in individual tumors of PCAWG (red) and TCGA (blue) datasets. **(E)** Pan-cancer distribution of the enrichment scores for CNA burden, CNA count, INDEL count and SBS count in PCAWG dataset. The enrichment scores are calculated as the ratio of mean value of a specific cancer type compared with the mean value of the whole PCAWG dataset. The colors indicate the false discovery rate (FDR) corrected P-values of Mann-Whiney U-test.

have been named as CNS1, CNS2...CNS14 throughout this study (Supplementary Figure 4). With the building of pan-cancer CN signature repertoire, for any cancer patient with absolute CN profile available, we can re-construct the composition of CN signature through single sample signature fitting (Supplementary Figures 5 and 6). Compared with our previous method [18], the new method reported here could reveal potentially unknown patterns, additionally it has the following advantages: (i) the uniqueness of signature profile reflected in the similarity comparison of CNA signature profiles is improved (Figure 3A). (ii) The reconstruction error in signature extraction is decreased (Figure 3B). Macintyre *et al.* applied a mixture modeling based method for CN component extraction, and the CNA component values are not consistent in different datasets, and this prohibits the signature comparison using cosine similarity analysis across different datasets [16, 18]. These differences between the CNA signatures extracted with Macintyre method, Wang method and this study have been illustrated using PCAWG ovarian cancer dataset (Supplementary Figure 7).

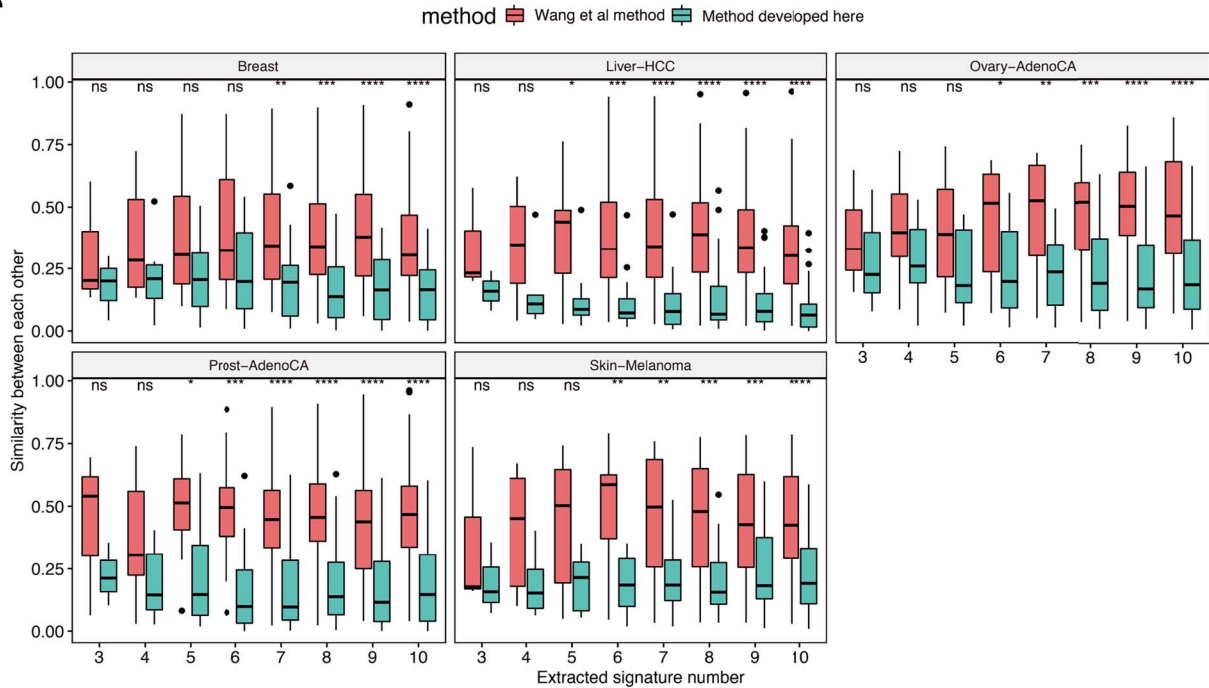
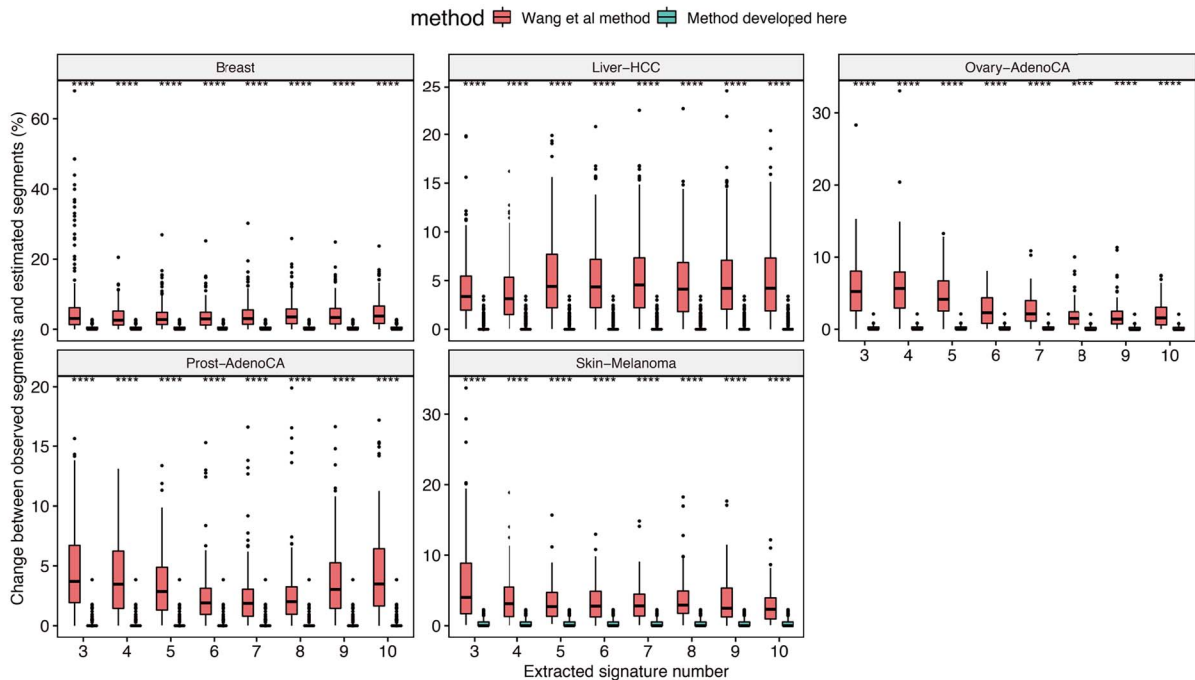
## Benchmark analysis of CNA signatures

To evaluate the robustness of the proposed CN signature analysis procedure, we compared the CNA signatures extracted in 2778 PCAWG WGS dataset with the CNA signatures derived from

10 851 TCGA SNP array dataset (Supplementary Figures 4 and 8). In 10 851 TCGA SNP array data, 20 CNA signatures have been extracted, these TCGA CNA signatures have been named as Sig1, Sig2...Sig20 throughout this study (Supplementary Figure 8). The similarities between signatures extracted from TCGA SNP array dataset and PCAWG WGS dataset are calculated using cosine similarity analysis method (Figure 4A). Most (9/14) PCAWG signatures can have highly similar (cosine similarity  $R \geq 0.8$ ) counterparts in TCGA dataset. Four PCAWG signatures (CNS4, CNS5, CNS6, CNS14) have intermediate similarity ( $R \geq 0.51$ ) counterparts in TCGA dataset. Majority of TCGA CNA signature (16/20) have median to high similar counterpart signature in PCAWG CNA signature set. Four TCGA CNA signatures (Sig15, Sig18, Sig9, Sig20) do not have matched PCAWG signature. All these four unmatched TCGA CNA signatures are likely tissue specific signatures. Using 10% relative activity as a cut off, TCGA Sig20 is only observed in TCGA testicular germ cell tumors (TGCT), and this type of cancer is not included in PCAWG dataset. TCGA Sig18 is only observed in TCGA thymoma cancer type, which is also not included in PCAWG dataset. TCGA Sig15 is majorly observed in adrenocortical carcinoma (ACC), and to a less extent in kidney chromophobe carcinoma, and ACC is not included in PCAWG dataset. TCGA Sig9 is observed in ACC and TGCT, both ACC and TGCT are not included in PCAWG dataset.



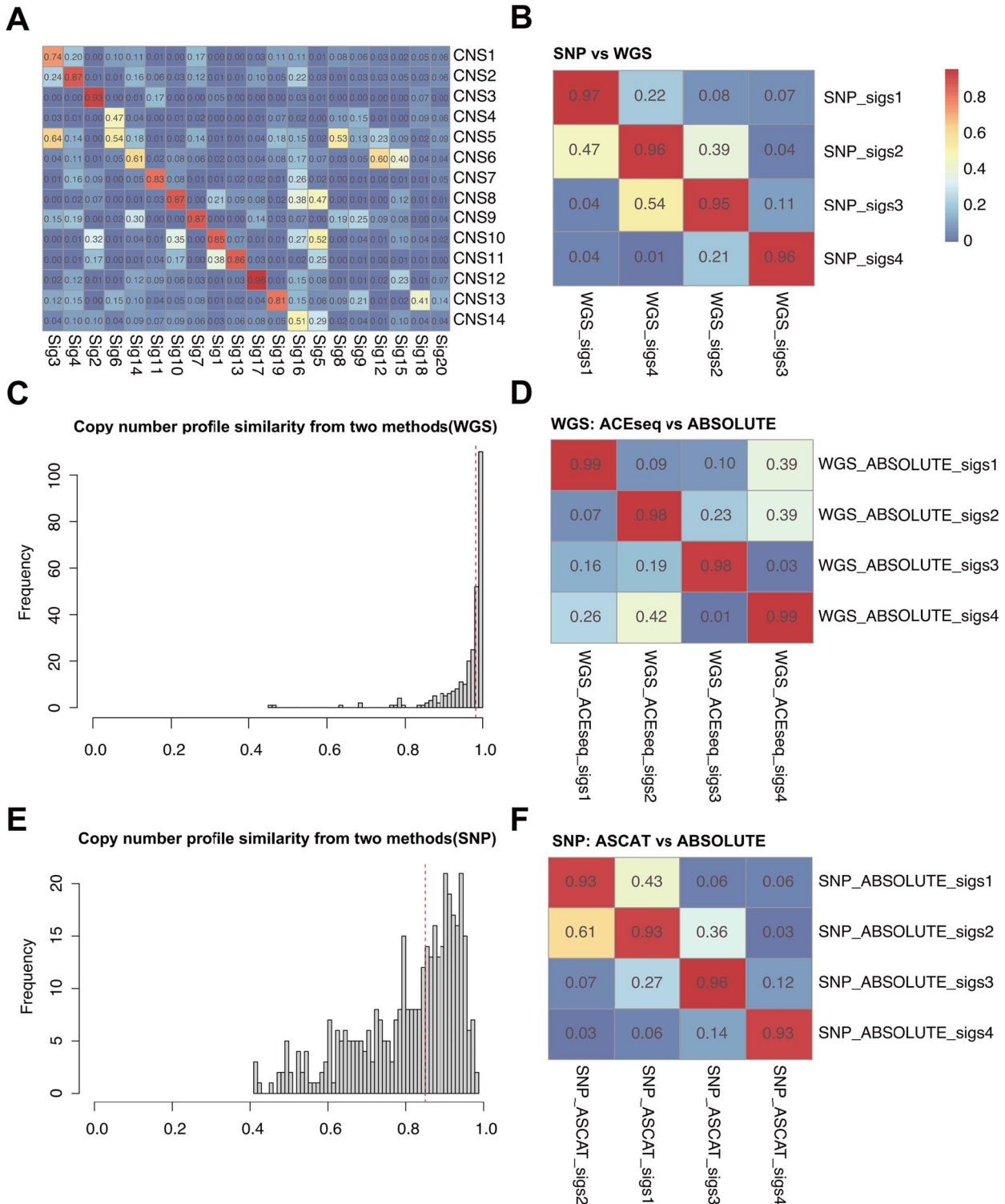
**Figure 2.** CNA classification strategy for signature analysis. For each CNA segment, the following features have been considered: (i) segment context, including segment shape and CN change number; (ii) Absolute CN; (iii) LOH status; (iv) segment size. In total, 176 types of CNA segments have been defined accordingly.

**A****B**

**Figure 3.** Comparisons between Wang method and the method reported here. **(A)** Inter-signature similarity comparison between Wang method and them method reported here. In five independent PCAWG cancer sub-datasets, consistently lower inter-signature similarities are observed with the method described here. **(B)** Comparison of signature reconstruction error between Wang method and the method reported here. The method reported here shows lower signature reconstruction error. Wilcoxon rank-sum test was performed to test for the differences between the two groups. ns:  $P > 0.05$ , \* $P \leq 0.05$ , \*\* $P \leq 0.01$ , \*\*\* $P \leq 0.001$ , \*\*\*\* $P \leq 0.0001$ .

The fact that CNA signatures extracted de novo from PCAWG (WGS based) and TCGA (SNP array based) show a pairwise similar signature profile suggests the robustness of our method in identifying true cancer patterns. To further evaluate the performance of our method, we carried out benchmark analysis with CNA

profiles derived from different platforms and different CNA calling algorithms. For benchmark analysis with CNA data from different platforms, CNA signatures have been extracted independently from 286 WGS and 468 SNP array derived prostate cancer CNA profiles. The CNA signatures extracted from WGS



**Figure 4.** CNA signature benchmark analysis with CNA profiles derived from different platforms and different CNA calling algorithms. **(A)** Inter-correlations between the CNA signatures extracted in PCAWG dataset and TCGA dataset. Cosine similarity values are reported for each comparison. **(B)** Cosine similarities are reported in comparing CNA signatures extracted from WGS platform and SNP array platform. **(C, D)** CNA profiles have been extracted from 286 WGS samples using ACEseq or ABSOLUTE algorithm, distribution of CN profile cosine similarities between the two methods is reported **(C)**, pairwise comparisons of the CNA signatures extracted with the CNA profiles called with the two algorithms are shown **(D)**. **(E, F)** CNA profiles have been extracted from 468 SNP array samples using ASCAT or ABSOLUTE algorithm, distribution of CN profile cosine similarities between the two methods is reported **(E)**, pairwise comparisons of the CNA signatures extracted with the CNA profiles called with the two algorithms are shown **(F)**.

data are highly similar to the CNA signatures extracted from SNP array data (median cosine similarity  $R=0.96$ ) (Figure 4B). The effects of different CNA calling algorithms on the stability of CNA signatures have been evaluated (Figure 4C and F). With WGS, the signatures derived from CNA profiles called with ABSOLUTE [25] algorithm are highly similar to the signatures derived from the CNA profile called with ACESeq algorithm (median cosine similarity  $R=0.98$ ) (Figure 4C and D). Similarly, for SNP array platform, the CNA signatures derived from ABSOLUTE algorithm are highly similar to the CNA profile derived from ASCAT [26] algorithm (median cosine similarity  $R=0.93$ ) (Figure 4E and F). In conclusion, the CNA signature extraction method proposed in this study can be applied to CNA profiles derived from different platforms and different CNA calling algorithms, and the CNA signatures extracted with our algorithm are stable and could reflect the true DNA alteration patterns of cancer.

### Pan-cancer distribution of CNA signature

The proportion of tumors with the signature and median activity of the signature in different types of cancers are shown for PCAWG dataset (Figure 5A). CNS3 was observed in tumors with few CNA counts, such as acute myeloid leukemia (AML) (Figure 5A and Supplementary Figure 9). The enrichment scores (defined as the ratio comparing the mean value of specific cancer type versus mean value of pan-cancer dataset) of the activities of CNA signatures in different types of cancer are shown for PCAWG dataset (Figure 5B). Some CNA signatures show enrichment in specific cancer type, for example, the enrichment score of CNS4 in PCAWG liposarcoma is 52.91 (Figure 5B and Supplementary Figure 10). The profile of CNS4 suggests the presence of extrachromosomal DNA (ecDNA) or neochromosome. Actually, double minutes, small, self-replicating extrachromosomal structures in a ring form, were originally observed in sarcomas [27]. The presence and evolution of neochromosome has also been investigated in liposarcomas [28]. Pan-cancer distributions of the relative and absolute activities of CNA signatures are shown for PCAWG dataset (Figure 5C and D). Compared to other signatures, CNS3 has the highest relative activities. Pan-cancer profiles of TCGA CNA signatures are shown (Supplementary Figure 11).

Compared with SBS, DBS, ID signature profiles, the profile of CNA signature showed much reduced inter-signature correlation, suggesting an increased signature distinctness compared with other types of signature profiles (Figure 6A and B). In PCAWG dataset, the average number of CNA signatures in a single patient is approximately two (Figure 6C). Some cancer types have few CNA signatures, such as AML. Some have many CNA signatures, such as breast cancer and ovary cancer (Figure 6D). This is in line with the fact that AML has a low number of CNA segment counts, while breast cancer and ovarian cancer have many CNA segment counts (Figure 1B). Several known CN patterns can be reproduced in our study using PCAWG dataset. For example: stable genome (CN-Sig 5 in Wang et al. 2021 study) [18]; ecDNA; chromothripsis; WGD, homologous recombination deficiency (HRD). Several new patterns have been identified in this study, including the following (Supplementary Figures 4 and 12): Pattern 1: Focal homozygous deletion (CNS10). This pattern is featured with regions of homozygous deletion. Generally homozygous deletion is focal and surrounding tumor suppressor genes. Pattern 2: Haploid chromosomes (CNS11). This pattern is characterized by a mixture of chromosomes with CN 1 and 2. This haploid status of several chromosomes could be caused by cell cycle defects. In the published literature, we can find that some cancer cells are haploid, and this could be derived through similar mechanism [29, 30]. Pattern 3:

Haploid chromosome and WGD (CNS6). This pattern is featured with chromosome LOH with CN 2 or 3. This pattern could be formed through haploid chromosome then WGD.

### Potential mutational processes for CNA signatures

An important purpose of CNA signature analysis is to identify the underlying mutational processes for CNA. Potential mutational processes for CNA include intrinsic inducers and extrinsic inducers. Intrinsic CNA inducers include: double-strand break repair defects (HRD, etc.); cell cycle defects; DNA replication defects; telomere loss, etc. Extrinsic CNA inducers include chemical or physical agents that induce double-strand breaks, or interfere with the cell cycle, such as chemotherapy drugs or ionizing radiation. Smoking and ultraviolet (UV) are known to induce SBS signatures, however, no specific CNA signatures are associated with smoking and UV (Supplementary Figures 13 and 14), probably because smoking and UV induce single-strand lesions, and are not associated with double-strand break, and the consequent CNA.

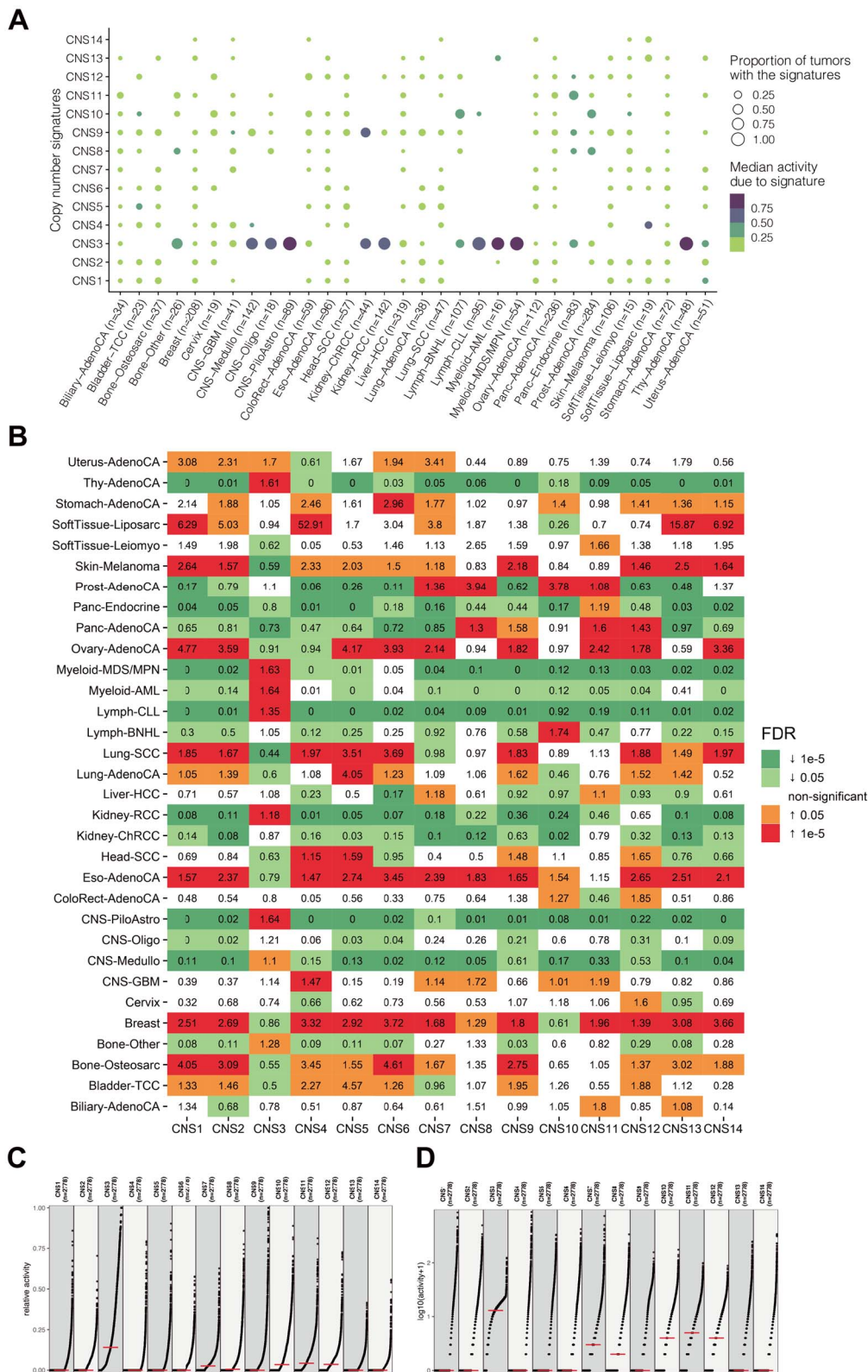
We group annotations of pathogenic germline variants and somatic driver mutations in DNA-repair genes across the PCAWG dataset, then correlate their presence with activities of the CN signatures (Figure 7A). BRCA1 functional mutations are significantly enriched in CNS14, suggesting the presence of HRD. CNS3 is significantly associated with patients without TP53 mutation, suggesting a state of stable genome. COSMIC SBS3 and ID6 are known HRD associated signatures, they also show strong correlations with CNS14 (Figure 7B). Associations between the presence of cancer driver mutations with the activities of CN signatures are shown for PCAWG and TCGA dataset (Supplementary Figures 15 and 16). Some driver mutations are significantly enriched in specific CNA signature. For example, SPOP mutation is specifically associated with CNS8 in prostate cancer (Supplementary Figures 15 and 17).

Correlations between some cancer genome features, such as CNA burden and the presence of ecDNA, with the activities of CNA signatures are displayed for PCAWG dataset (Figure 7C). In association analysis, activity of CNS9 is the most accurate predictor for WGD, suggesting CNS9 could be a reflection of the status of WGD. CNS14 activity is the most accurate predictor for HRD, suggesting CNS14 could be the major signature of HRD (Supplementary Figure 18). The activities of most CNA signatures show significant tumor stage difference, for example, CNS4 is present in higher level in stage III than in stage I tumors (Supplementary Figure 19). Representative sample profile, notable features and potential mechanisms or mutational processes for PCAWG and TCGA CNA signatures are summarized, respectively (Supplementary Figures 12 and 20). The mechanisms for several CNA signatures are still unknown.

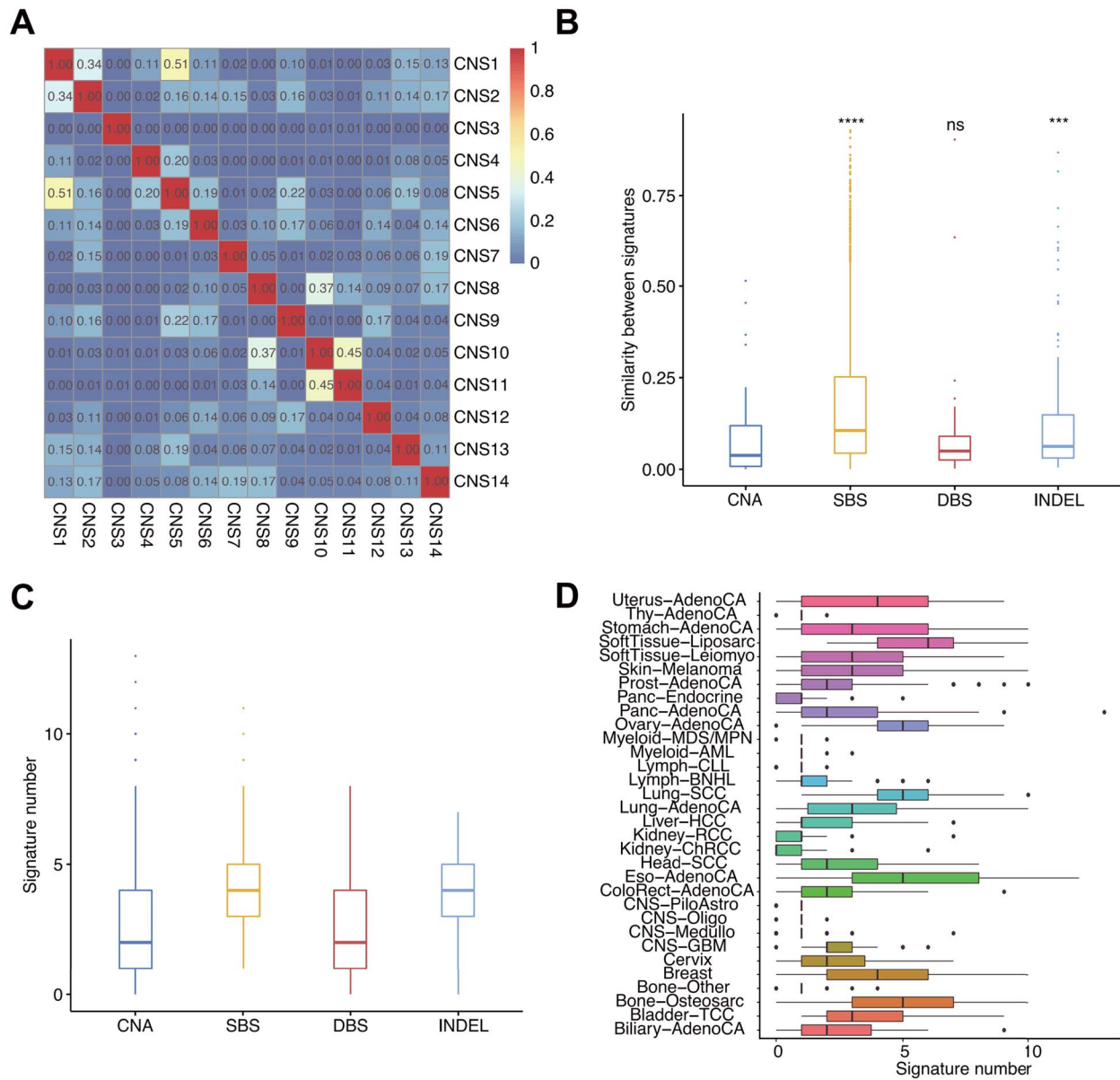
### The clinical relevance of CNA signatures

The CNA signatures extracted from cancer patients could be cancer prognosis biomarkers. To test this hypothesis, Cox regression analyses were conducted to evaluate the associations between the activity of each CNA signature and cancer patients' overall survival time for each cancer type. For each Cox model, we report a Z-score that encodes the directionality and significance of the survival relationship. A Z-score of  $>1.96$  indicates that the upregulation of the target feature (activity of CNA signature) is related to the reduction of the survival time at the  $P < 0.05$  threshold, while the Z-score of  $<-1.96$  indicates that the increase in the target feature at the  $P < 0.05$  threshold will indicate a longer survival time.





**Figure 5.** The activity distribution of CNA signatures in PCAWG pan-cancer dataset. **(A)** Proportion of tumors with the signature and the median activity of the signature are shown for 32 PCAWG cancer types. For each individual tumor, only signatures that contribute to  $\geq 5\%$  of the total are shown. **(B)** Enrichment score analysis of CNA signature in PCAWG dataset. The enrichment score is calculated as the ratio of mean signature activity of a specific cancer type compared with the mean signature activity of the whole PCAWG dataset. The colors indicate the FDR corrected P-values of Mann-Whitney U-test. **(C, D)** Relative activity (percentage of the total) **(C)** and absolute activity (contributed CNA segment number) **(D)** distribution pattern of CNA signatures in PCAWG dataset are shown.



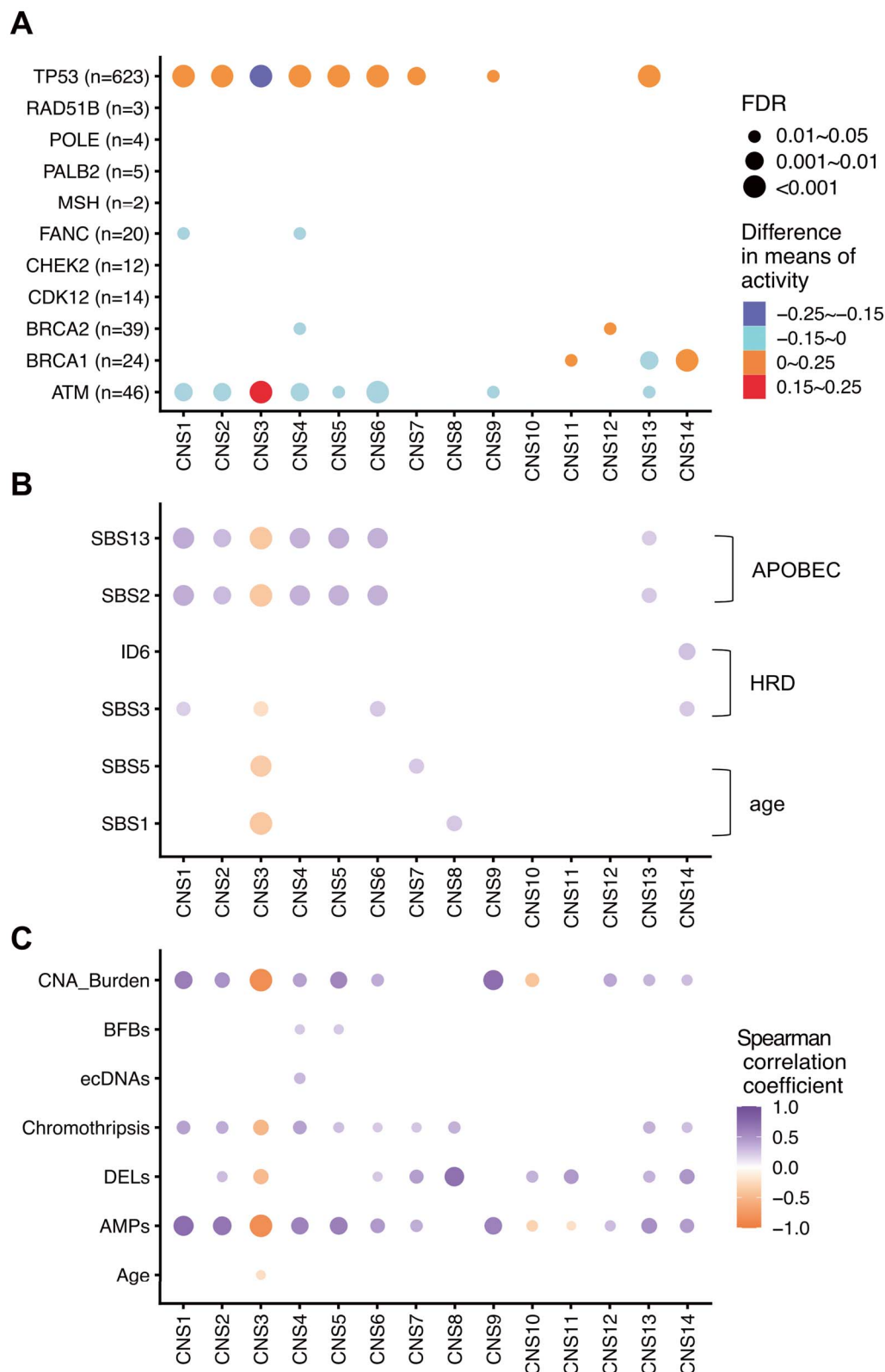
**Figure 6.** Distinctness and the number of CNA signatures extracted in PCAWG dataset. **(A)** Inter-correlation analysis of the profiles of 14 PCAWG CNA signatures. The numbers are cosine similarity values comparing each pair of CNA signatures. **(B)** Inter-signature similarities comparing the signatures of CNA, COSMIC reference SBS, DBS and ID (INDEL) signatures. **(C)** Median number of signatures for CNA, SBS, DBS, INDEL in PCAWG dataset. **(D)** Distribution of the number of CNA signature in each cancer type of PCAWG dataset.

To compare the prognostic effects of CNA signature activity in TCGA and PCAWG dataset, we select the cancer types that have sufficient number of patients ( $n > 50$ ) with both CNA signature activity and overall survival (OS) data available in both TCGA and PCAWG datasets. We calculate the CNA signature activity of both TCGA and PCAWG datasets using single sample fitting with TCGA signature set (Sig1, Sig2...Sig20). In total kidney cancer, stomach cancer, liver cancer, ovarian cancer and melanoma sub-datasets are available for this prognosis comparison analysis. Z-scores of CNA signatures are compared in pair in matching TCGA and PCAWG cancer types (Figure 8A). The Z-scores of CNA signature activity in matching TCGA and PCAWG cancer types are significantly correlated (Spearman correlation  $R = 0.44$ ,  $P = 0.015$ ), suggesting the robustness of CNA signature in predicting cancer patients' overall survival (Figure 8B). For example, in kidney cancer, high CNA Sig1 activity is associated with

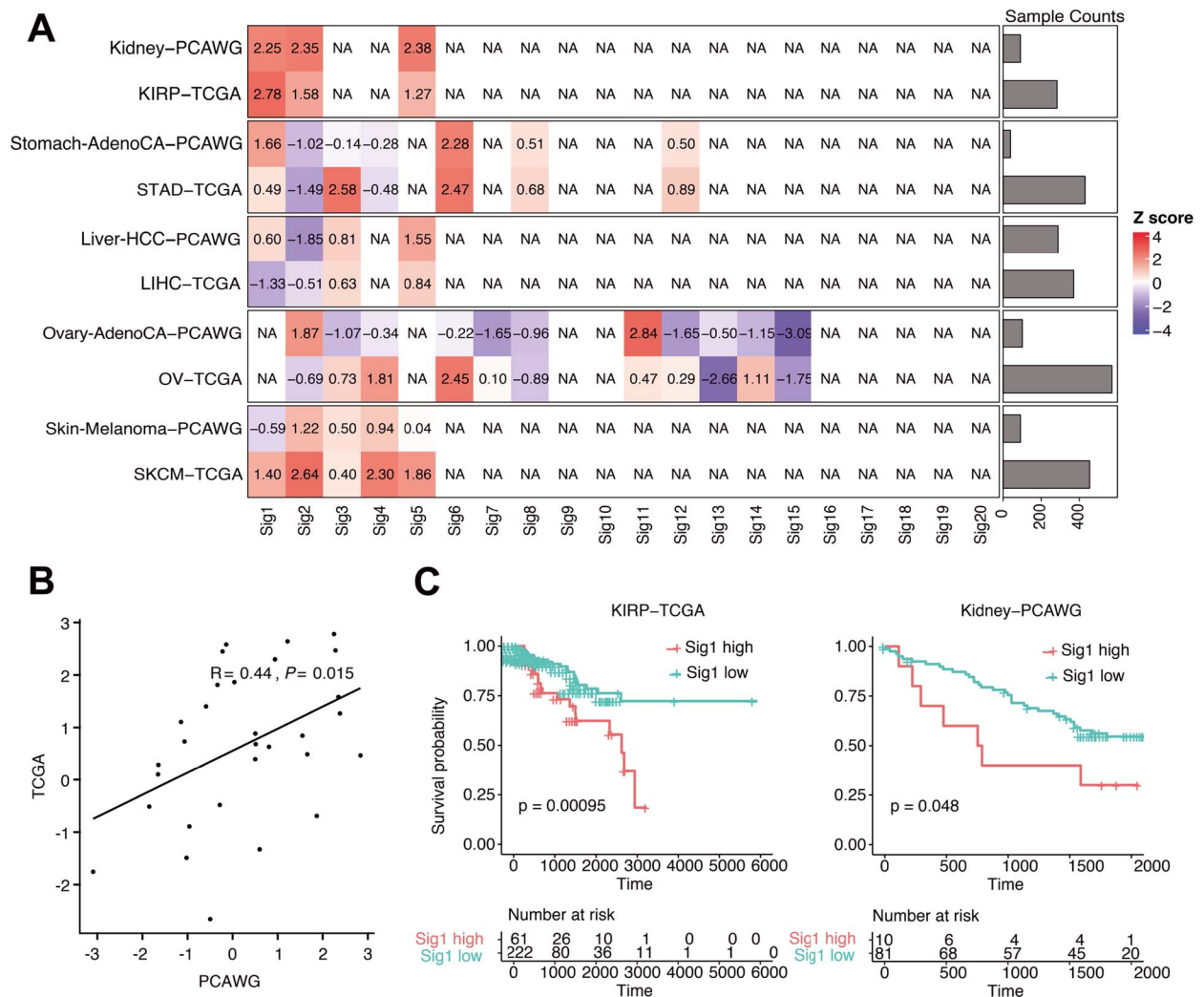
significantly poor overall survival in both TCGA and PCAWG datasets (Figure 8C).

## Discussion

Here we developed a unified and comprehensive method for CN signature analysis. Our method can be applied to cancer patients with CN profiles generated with WGS or SNP array data. Our CN signature analysis method is based on a novel and comprehensive method to catalog CN segments. New CNA patterns have been identified, and the activities of some CNA signatures have been demonstrated to be associated with cancer patients' prognosis. It will be of interest to investigate the potential application of CNA signatures in predicting the clinical response of certain cancer treatments, for example, PARP inhibitor treatment or cancer immunotherapy.



**Figure 7.** Correlations between CNA signatures and different types of genome alterations. **(A)** Associations between gene mutation status in key DNA-repair genes with the relative activities of CNA signatures. Association of pathogenic mutations (germline and somatic combined) in key DNA-repair genes with the relative activities of CNA signatures. For each gene, PCAWG patients are divided into two groups based on the mutation status of the gene. The difference values in mean relative activities between two groups (activities of mutated group—activities of un-mutated group) are reported, and FDR corrected *P*-values (Wilcoxon rank-sum test) are shown. The color and size of the points represent the differences and adjusted *P*-values, respectively. MSH refers to MSH2, MSH3, MSH4 and MSH6, genes in the mismatch repair pathway; FANC refers to genes associated with Fanconi anemia, namely FANCA, FANCC, FANCD2, FANCE, FANCF, FANCG, FANCI, FANCL and FANCM. The number after gene or gene cluster name refers to the number of pathogenic mutations (germline and somatic combined). **(B)** Inter-correlations between the activities of CNA signature and APOBEC, HRD, age signatures in PCAWG dataset. **(C)** Associations between the relative activities of each signature and the indicated cancer genome features are calculated from individual cancer patient. Spearman correlation coefficient values are reported.



**Figure 8.** CNA signature activity and cancer patients' prognosis. **(A)** The prognosis Z-scores of the activity of CNA signatures are calculated in each matching TCGA or PCAWG cancer type with sufficient number of patients ( $n > 50$ ) having overall survival and CNA profiles available for this analysis. NA indicates the CNA signature does not exist in the specific cancer type; colors reflect the values of Z-scores. **(B)** Correlations between the Z-scores of TCGA and PCAWG matching cancer types. Pearson test P-value is reported. **(C)** The activity of Sig1 shows consistent prognosis in the kidney cancer of TCGA (left) and PCAWG (right) datasets. Kaplan-Meier overall survival curves show the comparison between different groups stratified by CNA signature Sig1 activity. Samples with Sig1 activity higher than the cutoff (determined by `surv_cutpoint` function of 'survminer' package) were classified as 'Sig1 high' group, and samples with Sig1 activity less than the cutoff were classified as 'Sig1 low' group. The log-rank test P-values are reported. KIRP: Kidney renal papillary cell carcinoma; STAD: Stomach adenocarcinoma; LIHC: Liver hepatocellular carcinoma; OV: Ovarian serous cystadenocarcinoma; SKCM: Skin Cutaneous Melanoma.

Mutational signature analysis is initially developed with SBS, and CNA is different from SBS in several aspects. SBS and CNA are derived from different mutational processes, some common SBS inducers such as smoking and UV do not induce specific CNA signature. Smoking and UV can generate single-strand DNA lesions, while CNA is the consequence of DNA double-strand breaks (DSBs) or cell division defects [3]. The scale of the influenced genome is different between CNA and SBS. CNA could affect a much larger portion of the genome than SBS. Each SBS is usually a consequence of a unique mutational process, while each CNA segment could be the consequence of multiple CNA mutational processes. Some CNA inducers have an impact on global CNA pattern, such as inducers of cell division defects leading to WGD or aneuploidy, which have a global impact on CNA profile. The resulting signatures of global CNA mutational processes need to be combined with other CNA mutational processes, leading to new signatures. For example, WGD and HRD combination

generate new signature, which cannot be reconstructed as a linear combination of single WGD and single HRD signature. This suggested that the mutational processes for the CNA signatures reported in this study can be combinations of different processes. For example, CNS6 could be a combination of haploid chromosome (CNS11) and WGD (CNS9). The evolution timeline for mutational process of CNA is largely unknown. With the availability of cancer cell fraction (CCF) information for CNA segments [25, 31], we can calculate the activities of CNA signatures for clonal CNA and subclonal CNA, and this analysis can provide distinct insight into the evolution of CNA mutational processes and CNA patterns (Supplementary Figure 21).

Global patterns and mutational processes for CNA in human cancer are largely unknown. Currently known methods for CNA signature analysis include Macintyre method and Wang method, and both methods classify CNA segments using known patterns, such as the lengths of oscillating CN segment chains (named

'OsCN') [16, 18]. Some unknown CNA patterns that do not fit into these known patterns cannot be detected with these approaches. Here we provide a mechanism-agnostic method for CNA signature analysis, this method can reveal potentially new CNA patterns, and is suitable for pan-cancer study. For CN signature analysis, detailed sequencing information is not required. Compared with previously reported SV signature [15], the advantage of this method is the wider application area, for example, panel sequencing data or SNP array data could be used to derive CNA signature. The disadvantage is the increased uncertainty. Sometimes two nonequivalent genomes can produce exactly the same CN patterns. The underlying mutational process for CN pattern/signature could be complicated by these uncertainties.

During the submission process of this manuscript, Steele *et al.* reported pan-cancer CNA signature using a different mechanism-agnostic approach [17]. Compared to Steele *et al.* method, this study incorporates the context or shape information of each CNA segment, and this can reveal additional insight into the mutational processes and patterns of CNA. These CNA shape information include 'Low-Low', 'High-High', 'Ladder' and also the extent of CN change ( $>2$  or  $\leq 2$ ) (Figure 2). Some patterns of CNA can only be detected with our method but not Steele *et al.* method. For example, the oscillation status of CNA segment, the extent of CN change flanking specific CNA segment. The incorporation of this additional CNA segment shape or context information enables us to directly interpret the mechanism of CNA signature. Furthermore, our CNA signature profiles have been constructed and compared in two different pan-cancer datasets: TCGA SNP array dataset and PCAWG WGS dataset. The prognostic effects of CNA signatures have also been compared and validated in the above two different datasets, while Steele *et al.* did not perform signature profile and prognosis comparison analysis between TCGA and PCAWG datasets of different sample origin. Compared to recent publication, our study provides an innovative and comprehensive approach for CNA segment categorization and signature analysis, and reveals distinct insight into the patterns, evolution and mutational processes of CNA.

CNA signatures have just started to be investigated. Many issues still remain to be studied. One of the major purposes of CNA signature analysis is to reveal the underlying mutational process for CNA. The mutational processes for several CNA signatures reported in this study are unclear. This needs further association studies, and also experimental studies with pre-defined CNA inducers. Due to the heterogeneity of tumor, the final CNA profile can be a combined result of many heterozygous CNA difference. Single cell CNA signature analysis could reveal the heterogeneity and evolution process of cancer.

## Conclusions

Global patterns and mutational processes for CNA in human cancer are largely unknown. Here we developed a method to reveal pan-cancer patterns of DNA CNA signatures through a mechanism-agnostic approach. Our results correlate some CNA signatures to known biological characteristics through diverse approaches ranging from signature profile observation to molecular profiling. New CNA patterns have been identified, and the activities of some CNA signatures have been demonstrated to be associated with cancer patients' prognosis. Collectively, this method paves the way for further study on revealing new CNA etiology and designing robust biomarkers for cancer precision diagnosis and therapy.

## Materials and methods

### CN calling and processing

PCAWG allele-specific CN generated by PCAWG-11 working group is the result of a bespoke procedure that combines output from six different CN callers: ABSOLUTE, ACESeq, Battenberg, CloneHD, JaBbA and Sclust. First ran all methods across all samples with the consensus SVs included and applied an algorithm across the segmentations to obtain consensus breakpoints. With these mandatory breakpoints the methods were rerun without calling any additional breakpoints. TCGA allele-specific CN data were generated from Affymetrix Genome-Wide Human SNP 6.0 (SNP6) array with ASCAT2 workflow. The PCAWG and TCGA allele-specific CN data values are integers. CNAs were generally classified as such:

- Amplification: segment total CN  $> 2$ .
- Homozygous deletion: segment total CN = 0.
- LOH: segment with minor CN = 0 and total CN  $> 0$  and size  $> 10$  Kb.
- Deletion: segment total CN  $< 2$ .

### CN segment classification

CN segments were classified as normal, amplification or deletion. They were further classified by size of the segment [small (S),  $< 50$  Kb; middle (M), 50–500 Kb; large (L), 500 Kb–5 Mb; extreme large (E),  $> 5$  Mb]. They were then further classified by the context of a segment (the shape), i.e. how total CN of segment on its left/right side compared to it (HH, left CN Higher & right CN Higher; LL, left CN Lower & right CN Lower; HL, left CN Higher & right CN Lower; LH, left CN Lower & right CN Higher). LH and HL have the same shape, so they are combined into LD (Ladder like shape). The segments were further classified by checking if they harbor LOH or not, and their absolute CN. For LOH segments, the total CNs were classified into 1, 2LOH, 3+LOH; for non-LOH segments, the total CNs were classified into 0, 2, 3, 4, 5, 6, 7, 8, 9+. Finally, a total of 176 mutually exclusive categories (described as components in this study) were cataloged. CN profile of each sample can be inputted into the classification algorithm above and each component will be counted to generate an integer vector. For multiple samples, a component-by-sample matrix will be generated and used for CN signature discovery. The data import and classification procedure were implemented as functions 'read\_copynumber' and 'sig\_tally' in R package Sigminer (<https://cran.r-project.org/package=sigminer>).

### CN signature extraction

CN signatures were extracted from component-by-sample matrix with golden standard tool SigProfiler v1.0.17 (<https://github.com/AlexandrovLab/SigProfilerExtractor>) with default parameters. Briefly, this de novo signature extraction includes the following six steps: dimension reduction, resampling, NMF, iteration, clustering and evaluation [11]. For each NMF, the initialization was with random numbers, and iterations were performed for 10 000–1 000 000 times until stable results are obtained. This NMF process was repeated for 100 times with resampling data. Clustering the decomposition matrixes to identify the number of signatures from 2 to 30. The SigProfiler has been successfully applied to TCGA and PCAWG pan-cancer data for multiple mutation types including SBS, DBS and INDEL. Two key parameters for determining signature number, stability measured by average silhouette and the average Frobenius reconstruction error were obtained from the result of SigProfiler. We selected the signature

extraction solution as the maximum signature number that meets the following criteria:

1. No over fit.
2. Stability should be at least local maximal.
3. Mean cosine distance should be as small as possible.

Based on the rules, we selected 14 signatures for PCAWG CN data and 20 signatures for TCGA CN data. The profile and activity for each signature were obtained, accordingly.

## CNA signature benchmark analysis

We select data from prostate cancer to evaluate the reproducibility of our method with benchmark analysis. A 468 SNP-array data are derived from TCGA, we used the ASCAT [26] and the ABSOLUTE [25] algorithm to generate allele-specific CN. A 286 WGS data are derived from PCAWG and ACEseq [32], and the ABSOLUTE [25] algorithms were used to obtain absolute CN profiles. CN segments were categorized into 176 classes as described above in each sample. The cosine similarities between CNA profiles are calculated according to this 176-element vector. First, we benchmarked the impact of platforms on CNA signature analysis. Four CNA signatures are independently extracted from WGS, WES or SNP array derived CNA datasets with our method described above. Pairwise comparisons between the four CNA signatures are reported. Next, we benchmarked the impact of different CNA calling algorithms in CNA signature analysis. ACEseq and the ABSOLUTE algorithm were applied to obtain absolute CN profile from 286 WGS data independently, and four CNA signatures have been independently extracted, compared and cosine similarity values of pairwise comparisons are reported. Similarly, ASCAT and the ABSOLUTE algorithm were applied to obtain absolute CN profile from 468 SNP-array dataset independently, and four CNA signatures have been independently extracted and compared.

## CN signature labeling and matching

We sorted all CN signatures based on their total activities to all samples. PCAWG data analysis is the major focus of this study. For PCAWG, we named 14 CN signatures from CNS1 to CNS14. For TCGA, we firstly named 20 CN signatures from Sig1 to Sig20. We further added extra labels to their names for better comparing the CN signatures between PCAWG and TCGA by following the rules:

1. We classified signature similarity based on cosine similarity values into four levels:  $\geq 0.8$  (High, H),  $\geq 0.51$  &  $< 0.8$  (Intermediate, M) and  $< 0.51$  (unmatched).
2. The results were combined for a TCGA signature matched to two or more PCAWG signatures, i.e. TCGA Sig3 were matched to PCWG CNS1 and CNS5 both in middle level similarity, so the signature was labeled as 'Sig3-CNS1(M)/CNS5(M)'.
3. A postfix with the matched similarity rank was used if a PCWAG signature was matched to two or more TCGA signatures, i.e. TCGA Sig8 was labeled as 'Sig8-CNS5(M)\_3', here three means this signature was the 3rd signature matched to CNS5.

## Group enrichment analysis

To comprehensively show the enrichment of a variable across cancer types, inspired by enrichment analysis in Maftools, we designed and implemented the group enrichment analysis as functions 'group\_enrichment' and 'show\_group\_enrichment' of R

package Sigminer. To illustrate how this analysis works, here we use CNA burden analysis for PCAWG breast cancer as an example. Firstly, we divided all PCAWG samples into two categories: Breast and non-Breast. Then we compared the means of CNA burden with Wilcoxon rank-sum test and calculated the ratio of the means, i.e. 1.56 means the average CNA burden in Breast cancers is 56% higher in non-Breast cancers. The result of the statistical test indicates if this result was randomly obtained. We used heatmap to visualize the final results and distinguished the different results based on both the mean ratio and statistical test result:

- If the comparison result is statistically non-significant, white color was used to fill the heatmap cell.
- If the ratio is  $> 1$  and the comparison result is statistically significant, red color was used to fill the heatmap cell.
- If the ratio is  $< 1$  and the comparison result is statistically significant, green color was used to fill the heatmap cell.

## Association analysis

Associations between the activities of signatures, associations between signature activity and gene mutation status were performed using one of two procedures: (i) for a continuous association variable, Spearman correlation was performed; (ii) for a binary variable, patients were divided into two groups and a Wilcoxon rank-sum test was performed to test for differences in average activities of signatures between the two groups. Associations between each CNA signature and WGD or HRD status were performed using a two-sided Fisher's exact test. All reported P-values were FDR corrected. Only associations with both  $P \leq 0.05$  and odds ratio  $> 1$  were reported. Full correlation network for continuous variables was constructed using R package 'correlation' (<https://cran.r-project.org/package=correlation>).

## Survival analysis

Associations between CN signature activities with overall survival were identified using univariate Cox-proportional hazard models in each cancer type. For each Cox model, a Z-score that encodes the directionality and significance of the survival relationship is reported. Z-scores reflect the normalized deviations from the mean of a normal distribution, and these Z-scores are calculated following similar procedures as previously described [33]. Briefly, the Cox model is given by:

$$h(t, X) = h_0(t)e^{\beta X}$$

Where  $t$  is the overall survival time,  $h(t, X)$  is the hazard function,  $h_0(t)$  is the baseline hazard.  $X$  is a potential prognostic variable. Z-score is calculated by dividing the regression coefficient  $\beta$  by its standard error.

The prognosis effects of the activity of CNA signatures are compared in matching TCGA and PCAWG cancer types with sufficient number ( $n > 50$ ) of patients having both CNA profile and overall survival data available for analysis. Here, the absolute activity of CNA signature was used. For mutation data, the absolute activity is explained as the estimated mutation count contributed by a signature; similarly, for CN data, the absolute activity is explained as the estimated segment count contributed by a signature. Kaplan–Meier survival analysis was performed using the R package 'survival' with log-rank test, and Cox-proportional hazard analysis was performed using the R package 'ezcox'.

The cutoff value in Kaplan–Meier overall survival analysis was determined by `surv_cutpoint` function of ‘`survminer`’ package.

### CNA signature evolution analysis

The estimated CCF value of each segment is derived from ABSOLUTE [25] algorithm. Firstly, we sort CNA segments by their estimated CCF in a given sample. Next, we infer an evolutionary trajectory of the CNA signature activities over the estimated ordering of the CNA segments. We convert the CNA segments ordering into a set of CCF cut points with overlapping subsets of segments by decreasing CCF. The absolute CNA signature activities (counts of CNA segments) of the CNA segment subset are calculated based on single sample CNA signature fitting with TCGA CNA signature set using quadratic programming.

### Statistical analysis

Data between two groups were compared using a two-tailed unpaired Student’s t-test or Wilcoxon rank-sum test (also known as ‘Mann–Whitney’ test) depending on normality of data distribution (Typically, preprocessed expression data are normally distributed and mutation data show non-normal distribution). Correlation analysis was performed using the Spearman method. All reported *P*-values are two-tailed, and for all analyses,  $P \leq 0.05$  is considered statistically significant, unless otherwise specified. Multiple testing *P*-values were corrected by Benjamini–Hochberg FDR method. ‘ns’ for non-significant ( $P > 0.05$ ); ‘\*’ for  $P \leq 0.05$ ; ‘\*\*’ for  $P \leq 0.01$ ; ‘\*\*\*’ for  $P \leq 0.001$ ; ‘\*\*\*\*’ for  $P \leq 0.0001$ . All statistical analysis was performed using R v4.3.

#### Key Points

- A mechanism-agnostic method for CNA classification and signature analysis has been constructed, this method can reveal unknown patterns of CNA compared with existing methods.
- This method achieves robust and consistent results in pan-cancer CN signature analysis compared with known methods.
- Pan-cancer patterns and mutational processes for CNA signatures have been revealed through our method.
- CN signature activity consistently predicts the prognosis of cancer patients.

### Supplementary data

Supplementary data are available online at <https://academic.oup.com/bib>.

### Acknowledgments

We thank ShanghaiTech University High Performance Computing Public Service Platform for computing services. We thank R.S. for editing the text. We thank multi-omics facility, molecular and cell biology core facility of ShanghaiTech University for technical help.

### Funding

Shanghai Science and Technology Commission (21ZR1442400), National Natural Science Foundation of China (31771373) and startup funding from ShanghaiTech University.

### Data availability

PCAWG patients’ data including allele-specific CN, tumor purity, tumor ploidy, tumor WGD status and general phenotype were obtained from <https://pcawg.xenahubs.net> [20, 34]. PCAWG gene expression data in count format were obtained from ICGC data portal (<https://dcc.icgc.org/releases/PCAWG>). Chromothripsis results detected by ShatterSeek were obtained from <http://compbio.med.harvard.edu/chromothripsis/>. PCAWG amplicons (including ecDNAs) detected by AmpliconArchitect were obtained from Kim et al. study [35, 36]. PCAWG HRD status were obtained from Nguyen et al. study [37]. PCAWG telomere contents detected by TelomereHunter were obtained from PCAWG-SV working group’s study [24]. PCAWG mutational signatures (including SBS, DBS and INDEL) detected by SigProfiler were obtained from Alexandrov et al. study [12]. TCGA allele-specific CN and gene expression in count format were obtained from GDC portal (<https://portal.gdc.cancer.gov/>). TCGA clinical data were obtained from <https://pancanatlas.xenahubs.net>.

### Code availability

All code required to reproduce the analysis outlined in this manuscript are freely available at [https://github.com/XSLiuLab/Pan-cancer\\_CNA\\_signature](https://github.com/XSLiuLab/Pan-cancer_CNA_signature). Analyses can be read online at [https://xsluolab.github.io/Pan-cancer\\_CNA\\_signature](https://xsluolab.github.io/Pan-cancer_CNA_signature).

### Contributions

ZT, SW, CW, collected the data, developed the CNA signature analysis method and performed the computational analysis. TW, XZ, WN, GW, JW, JC, KD, FC participated in critical project discussion and resources. XSL conceptualized the idea, designed, supervised the study and wrote the manuscript.

### References

1. Zack TI, Schumacher SE, Carter SL, et al. Pan-cancer patterns of somatic copy number alteration. *Nat Genet* 2013;**45**:1134–40.
2. Beroukheim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *Nature* 2010;**463**:899–905.
3. Yi K, Ju YS. Patterns and mechanisms of structural variations in human cancer. *Exp Mol Med* 2018;**50**:1–11.
4. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 2009;**458**:719–24.
5. Stephens PJ, McBride DJ, Lin ML, et al. Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature* 2009;**462**:1005–10.
6. Stopsack KH, Whittaker CA, Gerke TA, et al. Aneuploidy drives lethal progression in prostate cancer. *Proc Natl Acad Sci USA* 2019;**116**:11390–5.
7. Ben-David U, Amon A. Context is everything: aneuploidy in cancer. *Nat Rev Genet* 2020;**21**:44–62.
8. Li RY, Du YQ, Chen ZH, et al. Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science* 2020;**370**:82–9.
9. Hieronymus H, Murali R, Tin A, et al. Tumor copy number alteration burden is a pan-cancer prognostic factor associated with recurrence and death. *Elife* 2018;**7**:e37294.
10. Hieronymus H, Schultz N, Gopalan A, et al. Copy number alteration burden predicts prostate cancer relapse. *Proc Natl Acad Sci USA* 2014;**111**:11139–44.

11. Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;**500**:415–21.
12. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature* 2020;**578**:94–101.
13. Gulhan DC, Lee JJ, Melloni GEM, et al. Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat Genet* 2019;**51**:912–9.
14. Wang S, Jia M, He Z, et al. APOBEC3B and APOBEC mutational signature as potential predictive markers for immunotherapy response in non-small cell lung cancer. *Oncogene* 2018;**37**:3924–36.
15. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;**534**:47–54.
16. Macintyre G, Goranova TE, De Silva D, et al. Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet* 2018;**50**:1262–70.
17. Steele CD, Abbasi A, Islam SMA, et al. Signatures of copy number alterations in human cancer. *Nature* 2022;**606**:984–91.
18. Wang SX, Li HM, Song MF, et al. Copy number signature analysis tool and its application in prostate cancer reveals distinct mutational processes and clinical outcomes. *PLoS Genet* 2021;**17**:e1009557.
19. Wang SX, Tao ZY, Wu T, et al. Sigflow: an automated and comprehensive pipeline for cancer genome mutational signature analysis. *Bioinformatics* 2021;**37**:1590–2.
20. Wang SX, Xiong Y, Zhao LF, et al. UCSCXenaShiny: an R/CRAN package for interactive analysis of UCSC Xena data. *Bioinformatics* 2022;**38**:527–9.
21. Korbelt JO, Campbell PJ. Criteria for inference of chromothripsis in cancer genomes. *Cell* 2013;**152**:1226–36.
22. Menghi F, Inaki K, Woo XY, et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc Natl Acad Sci USA* 2016;**113**:E2373–82.
23. Consortium ITP-CAoWG. Pan-cancer analysis of whole genomes. *Nature* 2020;**578**:82–93.
24. Li Y, Roberts ND, Wala JA, et al. Patterns of somatic structural variation in human cancer genomes. *Nature* 2020;**578**:112–21.
25. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol* 2012;**30**:413–21.
26. Van Loo P, Nordgard SH, Lingjaerde OC, et al. Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci USA* 2010;**107**:16910–5.
27. Buoen LC, Brand KG. Double-minute chromosomes in plastic film-induced sarcomas in mice. *Naturwissenschaften* 1968;**55**:135–6.
28. Garsed DW, Marshall OJ, Corbin VDA, et al. The architecture and evolution of cancer neochromosomes. *Cancer Cell* 2014;**26**:653–67.
29. Drouin V, Viguie F, Debesse B. Near-haploid karyotype in a squamous-cell lung-carcinoma. *Genes Chromosomes Cancer* 1993;**7**:209–12.
30. Kotecki M, Reddy PS, Cochran BH. Isolation and characterization of a near-haploid human cell line. *Exp Cell Res* 1999;**252**:273–80.
31. Ha G, Roth A, Khattra J, et al. TITAN: inference of copy number architectures, in clonal cell populations from tumor whole-genome sequence data. *Genome Res* 2014;**24**:1881–93.
32. Kleinheinz K, Bludau I, Hübschmann D, et al. ACEseq – allele specific copy number estimation from whole genome sequencing. *bioRxiv*. 2017; 210807.
33. Smith JC, Sheltzer JM. Genome-wide identification and analysis of prognostic features in human cancers. *Cell Rep* 2022;**38**:110569.
34. Wang S, Liu X. The UCSCXenaTools R package: a toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-Seq. *J Open Source Softw* 2019;**4**(40):1627. <https://doi.org/10.21105/joss.01627>.
35. Kim H, Nguyen NP, Turner K, et al. Extrachromosomal DNA is associated with oncogene amplification and poor outcome across multiple cancers. *Nat Genet* 2020;**52**:891–7.
36. Wu T, Wu CX, Zhao XY, et al. Extrachromosomal DNA formation enables tumor immune escape potentially through regulating antigen presentation gene expression. *Sci Rep* 2022;**12**(1):3590.
37. Nguyen L, Martens JWM, Van Hoeck A, et al. Pan-cancer landscape of homologous recombination deficiency. *Nat Commun* 2020;**11**:5584.