Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

TFIIH protein imbalance does not predict exome mutation burden or cisplatin sensitivity across cancer models

Achab-Awelana Ernest

achabawelana@gmail.com

Kwame Nkrumah University of Science and Technology https://orcid.org/0000-0001-8411-7895

Research Article

Keywords: TFIIH, nucleotide excision repair, proteogenomics, cisplatin, protein complex stoichiometry, DNA damage response

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11738486/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Protein abundance can diverge from transcript levels, potentially masking DNA-repair defects. We assessed whether transcription factor IIH (TFIIH) core-subunit imbalance provides a biomarker of nucleotide excision repair-associated phenotypes in cancer.

Methods

In nine CPTAC tumor cohorts, core-subunit protein residuals were estimated after adjustment for cognate RNA, copy number and tumor purity, and benchmarked against eight unrelated complexes. Seven cohorts (698 tumors) contributed exome single-nucleotide variant (SNV) burden. A parallel CCLE/DepMap analysis scored 361 cell lines; cisplatin-response data were available for 287 GDSC1 and 243 GDSC2 lines. Models adjusted for cell lineage, proliferation, mean TFIIH RNA and MLH1 expression. Published functional repair classifications were matched in an exploratory five-line pilot.

Results

TFIIH residual co-regulation was lower than the control-complex mean in all nine tumor cohorts (mean Spearman correlation 0.379 vs 0.487; paired $p = 0.0039$). Imbalance did not associate with SNV burden (pooled standardized $\beta = -0.009$, 95% CI − 0.085 to 0.067) or improve held-out prediction ($\Delta R^2 = -0.0064$). Cisplatin association estimates were − 0.026 (95% CI − 0.149 to 0.097) in GDSC1 and + 0.007 (95% CI − 0.147 to 0.162) in GDSC2. No association survived false-discovery control across 606 exploratory drug tests; SLFN11 expression showed the expected cisplatin association. The functional pilot was too small for confirmatory validation.

Conclusions

Bulk TFIIH residual imbalance did not provide a reproducible biomarker signal for the indirect endpoints examined. Direct nucleotide excision repair assays are needed to establish whether other forms of TFIIH disruption track functional repair deficiency.

1. Introduction

Nucleotide excision repair (NER) removes helix-distorting DNA lesions through global-genome and transcription-coupled recognition pathways. The TFIIH machinery links DNA unwinding during lesion verification with RNA polymerase II transcription initiation, placing its activity at a biological intersection that is relevant to both carcinogenesis and the response to DNA-damaging therapy [1, 2]. The same dual

functionality makes a putative TFIIH-based biomarker difficult to interpret: altered protein abundance may reflect transcriptional demands or cellular composition without indicating defective excision repair.

A biomarker based exclusively on canonical DNA-repair gene mutation can miss non-genetic dysfunction. Protein-complex subunits are not independently stable: protein synthesis, assembly, and turnover jointly constrain stoichiometry [3, 4]. Mass-spectrometry-based studies have documented protein–RNA discordance and buffering of copy-number changes across cancer [5, 6]. Loss of stable TFIIH assembly is biologically plausible because germline disruption of individual components can destabilize the complex in human disease [7]. These observations motivate testing a complex-level measurement rather than a single gene.

Yet a plausible mechanism is not evidence of clinical predictive value. A historical association between ERCC1 immunostaining and adjuvant cisplatin response in non-small-cell lung cancer [8] was subsequently challenged by isoform-specific measurement limitations and an unsuccessful antibody-based validation [9]. More recent NER-deficiency studies have demonstrated the value of lesion-associated mutational patterns in ERCC2-related bladder cancer [10]. Together, these findings suggest that readout specificity and independent validation are indispensable when evaluating an indirect biomarker.

The testable question addressed here is narrower than whether TFIIH is biologically important for NER. We asked whether variability in the relative abundance of measurable TFIIH core subunits, after adjustment for each gene's transcript, copy number, and tumour purity, identifies an unusually coordinated protein-complex state and predicts exome SNV burden, cisplatin response, drug-response patterns, or repair-gene dependencies. The study explicitly compares TFIIH with other complexes and includes a known cisplatin-response marker, SLFN11, as a positive-control exposure. Inference is restricted to the measured bulk-proteomic score and available indirect endpoints.

2. Materials and methods

2.1. Data resources and scope

The reanalysis used archived processed matrices from CPTAC (tfiih_export_patched.zip and tfiih_controls.zip), CCLE proteomics, DepMap 24Q4, and GDSC fitted drug-response resources (celllines_export.zip). The tables contain gene-level protein, RNA, copy-number and tumor-purity features, cell-line identifiers and fitted drug responses. These are processed public-data extracts rather than newly generated biological measurements. The core calculations were reconstructed in independent Python scripts, which do not reproduce every implementation detail of the initial exploratory workflow. Upstream source resources are described in [11–14].

Tumour cohorts comprised breast carcinoma (BRCA), clear-cell renal carcinoma (CCRCC), glioblastoma (GBM), head and neck squamous carcinoma (HNSCC), lung squamous carcinoma (LSCC), lung adenocarcinoma (LUAD), ovarian carcinoma (OV), pancreatic ductal adenocarcinoma (PDAC) and uterine

corpus endometrial carcinoma (UCEC). COAD was excluded from the pooled primary analyses because its proteomic measurements came from a different processing pipeline. BRCA and CCRCC lacked the corresponding cohort-specific somatic mutation tables in the supplied primary set and were therefore retained for the protein-complex benchmark but omitted from the primary pooled burden association.

2.2. TFIIH annotation and missing-data criteria

The human TFIIH core contains ERCC2/XPD, ERCC3/XPB, GTF2H1/p62, GTF2H2/p44, GTF2H3/p34, GTF2H4/p52 and GTF2H5/TTDA; the associated CDK-activating kinase module contains CDK7, CCNH and MNAT1 [1, 7]. GTF2H2 was excluded from the reanalysis because the original extracts described ambiguous peptides shared with GTF2H2C and insufficient reliable protein quantification. The primary operational core therefore comprised six genes: ERCC2, ERCC3, GTF2H1, GTF2H3, GTF2H4 and GTF2H5. A core gene had to be quantified in at least 70% of the corresponding tumour cohort, and a sample needed observed values for at least k − 1 of the k eligible subunits. Ovarian tumours had five eligible proteins; the other eight cohorts had six. No missing protein values were imputed.

2.3. Post-transcriptional residual and imbalance scores

For each gene g within each tumour cohort, its measured protein abundance Pg was modelled by ordinary least squares as $P_g = \alpha_g + \beta_{1g} \cdot RNA_g + \beta_{2g} \cdot CN_g + \beta_{3g} \cdot purity + \varepsilon_g$. Predictors were standardized within cohort; complete-case observations were used for each gene. The standardized residual ε_g was interpreted as protein abundance beyond that explained by cognate RNA, copy number and purity, not as a direct measurement of protein turnover. For each eligible tumour, the main imbalance score was the sample standard deviation of observed standardized residuals across core genes; the depletion score was their mean. The RNA covariate was the mean standardized abundance of core transcripts. Proliferation was summarized from available MKI67, PCNA, TOP2A and MCM2 proteins.

For cell lines, the analogous gene-level model used protein ~ own RNA + own copy number + cell-lineage indicators. The gene-level protein table was constructed from the published CCLE TMT columns, averaged when the same model appeared in multiple plexes, and linked to DepMap model IDs through CCLE names. The corresponding lineage, RNA, copy-number and MLH1 expression data were obtained from the supplied DepMap extracts. TFIIH imbalance was scored only if at least five of the six core proteins had residuals. TMT plex was included in a sensitivity analysis of drug associations; it was not included in the original score model.

2.4. Control-complex specificity benchmark

Eight reference complexes were specified from the supplied complexes.json list: MCM2–7 helicase, CCT/TRiC chaperonin, 20S proteasome, 19S proteasome, RNA polymerase II, TFIID, COP9 signalosome and RNA exosome. Each protein was residualized using the same gene-specific adjustment. The mean pairwise Spearman coefficient among residuals within each complex was calculated in each cohort. The TFIIH coefficient was compared with the within-cohort average across the eight controls, and cohort-

level differences were assessed with a two-sided Wilcoxon signed-rank test. This benchmark is a specificity control, not proof of physical protein–protein interaction.

2.5. Tumour mutation-burden association and prediction

The primary available genomic endpoint was the count of exome SNVs from the exported mutation files, rather than a clinically calibrated tumour mutational burden per megabase. The outcome was log10(nSNV + 1), rank-inverse-normal transformed within each cohort. Standardized cohort-specific OLS models adjusted for tumour purity, proliferation and mean core TFIIH RNA. HC3 heteroskedasticity-robust standard errors and 95% Wald intervals were used for the independent sensitivity reanalysis. Seven cohort-specific coefficients were combined by inverse-variance fixed-effect meta-analysis; a DerSimonian–Laird heterogeneity sensitivity calculation was also performed. The fixed-effect estimate is a summary of study-specific associations, not a causal population effect.

A leave-one-cancer-cohort-out prediction comparison contrasted a baseline model containing tumour purity, proliferation and mean TFIIH RNA with a model adding TFIIH imbalance and depletion. Predictors were standardized within cancer type to limit intercohort scale differences; linear models were fit on the other six cohorts and evaluated on the held-out cohort. Pooled held-out R^2 was used as a descriptive incremental prediction check. No external cohort was available for prospective confirmation.

2.6. Cell-line cisplatin association

GDSC1 and GDSC2 fitted cisplatin AUC values were matched to DepMap model IDs through Sanger model IDs, and repeated records within the same drug–model pair were averaged. AUC was standardized separately within each dataset; higher AUC indicates relatively less drug sensitivity within the dataset's assay definition. The primary explanatory variable was standardized TFIIH imbalance, with adjustment for lineage (rare categories combined), proliferation, mean TFIIH RNA and MLH1 RNA as an imperfect mismatch-repair proxy. Inference used HC3-robust errors. A sensitivity specification additionally adjusted for TMT plex. SLFN11 transcript abundance was tested with the same covariate structure as an assay and mapping positive control [15].

2.7. Exploratory drug and gene-dependency screens

The independent cross-check analysed each drug–dataset pair with at least 100 model-matched cell lines, yielding 606 comparisons across GDSC1 and GDSC2. Each model used the same lineage, proliferation, TFIIH RNA and MLH1 covariates as the cisplatin model. Benjamini–Hochberg correction was applied jointly to the 606 robust p values. A CRISPR sensitivity screen examined 17 available NER/TFIIH/BRCA-related gene effects in 302 cell lines, with the same main adjustment and Benjamini–Hochberg correction across the 17 tests. GTF2H2 dependency was not present in the supplied subset and was not silently imputed. These screens are explicitly exploratory.

2.8. Statistical interpretation, reproducibility and limitations of source extracts

The independently executable scripts reproduce the core analytical questions but differ from some preliminary calculations in variance estimation, cell-line lineage grouping, duplicate TMT-model aggregation and cohort-specific complete-case processing. The reported confidence intervals quantify uncertainty under the specified models and should not be interpreted as formal equivalence bounds. Neither bulk exome SNV burden nor broad substitution classes are direct NER activity measurements. The available exports lack trinucleotide context, transcriptional orientation and lesion-removal kinetics needed for pathway-specific interpretation [2, 10, 16]. No external prospective registration was performed; the control-complex benchmark and the functional-phenotype pilot are distinguished from the primary analyses.

AI-assisted scripting was used during preparation of portions of the Python analysis pipeline (OpenAI ChatGPT). Executable code and outputs were checked through an independent reanalysis of the available processed files. The use of AI assistance does not substitute for author responsibility for model specification, reproducibility or interpretation.

2.9. Exploratory triangulation against published functional NER phenotypes

As a post hoc biological triangulation analysis, we extracted cell-line level NER phenotypes from the experimental fluorescence/UV-based functional classification reported by Wei and colleagues [19] and the ERCC4-silenced MDA-MB-468 model documented by Rajkumar-Calkins and colleagues [18]. The six Wei et al. models comprise three NER-deficient (HCC1806, MCF7, SKBR3) and three NER-proficient (MDA-MB-231, MDA-MB-361, MDA-MB-453) models. We added MDA-MB-468 as an independently reported deficient model, without implying that the measurement platforms, thresholds or study selection were identical. Experimental NER activity was not newly measured in this work.

Names were mapped to the DepMap 24Q4 Model table by exact normalized cell-line names, then to TFIIH residual scores calculated from the supplied CCLE proteomics subset. Models without an evaluable score under the six-core-subunit minimum-completeness rule were excluded rather than imputed. Two classification sets were reported: the within-study Wei subset and a combined descriptive set. We tested the originally hypothesized direction (higher TFIIH residual imbalance indicating poorer repair); discrimination was summarized by area under the ROC curve and exact permutation enumeration of all class-label assignments. The analysis was planned after observing the original global null results, is exploratory, has no estimated clinical threshold, and is not a preregistered confirmatory analysis. We additionally extracted normalized ERCC4 and GTF2H1 protein signals and empirical percentiles from the CCLE proteomics subset, avoiding claims of absolute protein copy number.

3. Results

3.1. Cohort coverage and data integrity

Nine CPTAC cohorts contributed to the tumor-score framework (901 complete-case observations in available burden-model subsets), including 698 tumors from seven cohorts with cohort-specific mutation files. Six core proteins were evaluable in all cohorts except ovarian cancer, where five met coverage thresholds. HNSCC copy-number data were obtained from a distinct source and BRCA/CCRCC union mutation records were excluded from the primary analysis. The reconstructed cell-line pipeline yielded 361 scoreable models, of which 287 and 243 had cisplatin AUC measurements in GDSC1 and GDSC2, respectively. These GDSC subsets overlap in cell-line identity and must not be regarded as independent patient cohorts.

3.2. TFIIH co-regulation is detectable but not exceptional

Across nine cohorts, the mean pairwise Spearman correlation of TFIIH core residuals averaged 0.379, indicating positive residual coordination after gene-level RNA, copy number and purity adjustment. However, the eight control complexes averaged 0.487. TFIIH was below the cohort-specific control mean in every cohort (mean difference − 0.108; two-sided Wilcoxon signed-rank p = 0.0039; Fig. 1). Thus, a signal that appears coordinated against a gene-permutation null is not unusually strong in the context of other cellular protein complexes. The relative ordering does not establish TFIIH instability: differential peptide detectability, distinct assembly states and TMT batch factors remain plausible contributors.

3.3. No pooled association with exome SNV burden

Across GBM (n = 99), HNSCC (109), LSCC (108), LUAD (109), OV (78), PDAC (100) and UCEC (95), the pooled standardized imbalance–SNV association was β = −0.009 (approximate 95% CI − 0.085 to 0.067; fixed-effect inverse-variance pooling using HC3 errors). Between-cohort heterogeneity was small in this sensitivity model (I^2 approximately 7.5%). The ovarian cohort had the most negative coefficient (β = −0.271; 95% CI − 0.543 to 0.002) but did not cross the conventional two-sided threshold under robust inference, and this isolated estimate should not be interpreted as a validated ovarian-specific signal. BRCA and CCRCC were excluded from the primary estimate because their available exome records were assembled from a pan-cancer union file rather than the corresponding cohort-specific source.

3.4. TFIIH scores do not improve held-out burden prediction

Leave-one-cohort-out fitting of seven cancer types gave a held-out R^2 of 0.1097 for the baseline features and 0.1033 after adding TFIIH imbalance and depletion ($\Delta R^2 = -0.0064$). The direction argues against meaningful predictive gain in this cohort-transfer setting. Because the same source cohort collection was used for score construction, this is internal cross-cohort validation rather than external clinical validation.

3.5. Cisplatin null associations survive robust and batch-sensitive modelling

In 287 GDSC1 models, the robust standardized imbalance coefficient for cisplatin AUC was − 0.026 (95% CI − 0.149 to 0.097; p = 0.680). In 243 GDSC2 models, the coefficient was + 0.007 (95% CI − 0.147 to 0.162; p = 0.927). Additional TMT plex indicators did not reveal a positive result: the coefficients were + 0.009 (p = 0.914) and − 0.047 (p = 0.572). In contrast, increased SLFN11 expression was associated with lower cisplatin AUC (GDSC1 standardized $\beta = -0.300$, $p = 1.1 \times 10^{-5}$; GDSC2 $\beta = -0.426$, $p = 2 \times 10^{-6}$), supporting successful source linkage and sensitivity of the modelling pipeline to a known pharmacogenomic signal (Fig. 3).

3.6. Broad drug and CRISPR screens do not yield FDR-supported discoveries

Across 606 drug–dataset tests, 34 had nominal $p < 0.05$ but none survived Benjamini–Hochberg $q < 0.10$ with HC3-robust inference. The strongest nominal signal was GDSC2 dactinomycin (standardized $\beta = -0.227$; $p = 3.8 \times 10^{-4}$; $q = 0.232$), which remains hypothesis-generating. Among 17 available CRISPR gene-dependency outcomes, the smallest nominal p was 0.054 and none survived $q < 0.10$. Exploratory drug associations were sensitive to variance specification and multiplicity handling; these results should not be interpreted as independently replicated findings.

3.7. Quantitative summary

Table 1
Summary of independently reconstructed TFIIH biomarker analyses.

Endpoint	Independent estimate	Interpretation
Control-complex benchmark (9 cohorts)	TFIIH r = 0.379; control mean r = 0.487; p = 0.0039	Lower TFIIH residual coordination
Exome SNV burden (n = 698)	Pooled $\beta = -0.009$; 95% CI − 0.085 to 0.067	No pooled association detected
Held-out SNV burden prediction	$\Delta R^2 = -0.0064$	No incremental predictive value
Cisplatin, GDSC1 (n = 287)	$\beta = -0.026$; 95% CI − 0.149 to 0.097	No association detected
Cisplatin, GDSC2 (n = 243)	$\beta = +0.007$; 95% CI − 0.147 to 0.162	No association detected
Other drugs (606 tests)	0 tests at BH $q < 0.10$	No corrected discoveries
CRISPR dependencies (17 genes)	0 tests at BH $q < 0.10$	No corrected discoveries

3.8. Cross-study functional NER-phenotype triangulation (exploratory)

Five of seven externally classified models had evaluable TFIIH imbalance scores: four of the six lines examined by Wei et al. [19] and the ERCC4-silenced MDA-MB-468 line reported by Rajkumar-Calkins et al.

[18]. Three of these five were classified as functionally NER-deficient and two as NER-proficient. In the four-line within-study subset, TFIIH imbalance was 0.361 and 0.220 for the two deficient models (HCC1806, MCF7), compared with 0.814 and 0.688 for the proficient models (MDA-MB-231, MDA-MB-453). The additional deficient model MDA-MB-468 had imbalance 0.489. All five scores therefore ranked in the direction opposite to the proposed high-imbalance/NER-deficiency biomarker; the combined descriptive ROC AUC for the hypothesized direction was 0.00 (exact two-sided permutation $p = 0.20$; five models, ten possible class allocations). The within-study four-model AUC was also 0.00 (exact two-sided $p = 0.33$). These very small samples cannot establish inverse biology and do not permit reliable estimates of discrimination or clinical utility.

The MDA-MB-468 line, previously shown to possess ERCC4 methylation-associated XPF deficiency [18], exhibited an ERCC4 normalized protein measurement in approximately the third percentile of the available CCLE panel; its GTF2H1 protein value was around the 42nd percentile. The observations are consistent with an NER defect occurring outside TFIIH stoichiometric imbalance, but no causal inference can be made from this cross-platform observation. This analysis uses external published phenotype classifications rather than harmonized numerical repair kinetics and should be regarded as pilot evidence. Figure 4 and Supplementary Table S1 provide all matched model values and exclusions.

4. Discussion

4.1. Main inference: measurement of a complex is not measurement of repair

The principal result is a failure to detect a useful transferable association between bulk TFIIH residual stoichiometry and the indirect tumour or drug-response phenotypes examined. The framework was sufficiently informative to reveal strong positive residual coordination among core proteins, but that association was less pronounced than the average of unrelated protein complexes. This is the distinguishing methodological lesson: protein co-regulation that is statistically significant against a naive permutation baseline may still be biologically ordinary when a relevant reference distribution is considered.

Mechanistically, TFIIH can be disrupted by amino-acid substitutions, altered protein interactions, defective recruitment to damage, impaired helicase function, or separation of repair and transcription activities without substantially changing bulk protein abundances [1, 2, 7]. Conversely, cells may tolerate modest dispersion in steady-state protein concentrations while maintaining enough assembled holo-complex for NER. The measured coefficient and score cannot distinguish these states. It would therefore be scientifically incorrect to conclude that TFIIH is dispensable for NER or that functional repair deficiency is absent in all tested tumours.

4.2. Why the negative pharmacogenomic result is informative

The cisplatin result is notable because it was concordant across two GDSC assay subsets, survived robust heteroskedasticity adjustment, and was not converted into a positive signal by including TMT plex. The SLFN11 positive control showed a large association in the expected direction under the same identifier linkage and covariate pipeline, reducing concern that a completely broken cell-line merge explains the TFIIH null. Nevertheless, cisplatin AUC is influenced by growth rate, other repair pathways, drug transport, apoptosis and tissue lineage; a null in this surrogate cannot exclude subtle or context-restricted NER alterations. The two GDSC releases also include overlapping biological models, limiting their status as independent replication cohorts.

Apparent drug-screen signals depend on model specification and multiplicity handling. With HC3-robust uncertainty, no tested compound met the reported false-discovery threshold. In particular, the nominal dactinomycin association is compatible with a transcription-stress hypothesis but is not a corrected discovery and has not been independently replicated.

4.3. A meaningful negative result requires precise limits

The 95% cisplatin intervals are narrow enough to argue against a large linear association per standardized unit of the current score in these modelled cell-line sets. They are not formal evidence of equivalence because no clinically justified equivalence margin or pre-specified two one-sided equivalence testing framework was used. The finding is specific to the measured gene residual dispersion, bulk proteomic resolution and model covariates. It does not exclude nonlinear threshold effects, highly rare TFIIH-deficient states, cell-type mixtures obscuring single-cell perturbations or highly context-dependent vulnerabilities. Several subunits, especially GTF2H2, were inadequately measured, so the score cannot be interpreted as a complete holo-complex stoichiometric measurement.

4.4. The next experimental test must measure NER itself

Future high-impact validation should prioritize directly measured NER competence, not additional correlations with global exome mutation counts. Candidate readouts include UV-induced unscheduled DNA synthesis, cyclobutane pyrimidine dimer and 6 – 4 photoproduct clearance kinetics, damage-dependent recruitment of TFIIH, excision-product sequencing, and transcription-coupled lesion removal assays [2, 16]. In appropriately powered human tumour or cell-line collections, a genuine NER-specific mutational signature would require base-level context and ideally strand asymmetry; the six broad substitution fractions included here cannot provide that specificity. A nested design combining proteomics, targeted TFIIH perturbation and functional readouts would directly test the threshold of complex destabilization necessary to impair repair. Until those data exist, the score is best viewed as an exploratory measure of protein organization rather than a candidate clinical NER diagnostic.

4.5. Strengths and limitations

Strengths include nine molecularly annotated tumor cohorts, a residual approach conditioned on cognate RNA and copy number, an explicit non-TFIIH reference-complex benchmark, cross-cohort held-out assessment, two GDSC assay subsets and an informative SLFN11 positive control. The broad drug

screen reports multiplicity correction rather than selecting nominal results. Processed analytical matrices and independent Python scripts are available as part of the author's reproducibility materials.

Limitations are substantial. The initial exploratory scripts and complete time-stamped decision history were not available for one-to-one reproduction. Reanalysis used fixed gene panels instead of complete raw proteomic spectra, without original mass-spectrometry quality controls or tumor TMT-batch identifiers. TFIIH and reference proteins differ in detectability and subunit architecture. Mutation burden is an indirect phenotype influenced by multiple repair pathways and environmental exposures; lesion-context and strand-specific annotations were absent. GTF2H2 quantification remains ambiguous. CCLE/GDSC cell lines are not patient-treatment cohorts. The available DepMap extract lacked the GTF2H2 gene-dependency endpoint, so 17 rather than 18 targets were independently tested. The ovarian estimate, dactinomycin association and five-line functional comparison were exploratory, and the pooled seven-cohort estimate should not be generalized without qualification across histologies.

4.6. Implications of functional triangulation and mechanistic boundary

The pilot provides a functional anchor for the original null finding, but does not rescue the manuscript into a clinically validated biomarker study. In particular, repair loss owing to ERCC4/XPF deficiency [18] is biologically distinct from an inferred dispersion of TFIIH subunit residuals. The literature also already establishes a context-specific GTF2H1 mechanism: SWI/SNF ATPase perturbation can reduce GTF2H1 abundance and impair UV-damage responses, with compensatory restoration in some models [17]. These existing results rule out a novelty claim based simply on demonstrating GTF2H1 involvement in NER. A publishable mechanistic extension would require measuring lesion removal and TFIIH assembly within matched isogenic contexts, while contrasting GTF2H1 perturbation, a non-TFIIH repair factor such as ERCC4, and rescue conditions. This is a future experimental hypothesis, not a result of the present analysis.

The publicly archived GSE146473 XR-seq dataset includes ten colorectal cell lines subjected to oxaliplatin-induced lesion repair mapping [20]. Five of those lines could be unambiguously mapped to complete TFIIH scores in the supplied extracts (RKO, LS180, HCT116, HT29, SW480), but normalized repair-activity outcomes were not extracted and verified here; no corresponding XR-seq association is therefore claimed. Cross-study name overlaps alone cannot substitute for direct validation, particularly because transport-mediated platinum resistance and lesion formation can confound repair associations [20].

5. Conclusions

A coordinated bulk TFIIH residual-proteomics phenotype is measurable but not exceptional against other protein complexes, and it provided no incremental prediction of exome SNV burden or cisplatin AUC in the supplied cancer cohorts and cell-line collections. The interpretation is a bounded negative biomarker result, not evidence that TFIIH is functionally irrelevant or that repair capacity is invariably normal. Future

translation will require functional NER readouts, improved complex measurement and independent biological replication. More broadly, multi-protein biomarker discovery should use unrelated-complex controls, holdout assessment, working assay-positive controls and transparent multiplicity handling before claiming disease-specific signal.

Declarations

Data and code availability

The study reused processed extracts from publicly available CPTAC, CCLE, DepMap and GDSC resources [11–14] together with published functional NER classifications [18,19]. The processed input matrices, independent Python scripts, output tables, figures and five-line phenotype audit are supplied in an accompanying reproducibility package. This package does not contain original raw mass-spectrometry files or the earliest exploratory scripts. A permanent public repository accession and DOI are pending and must be added before submission. Functional labels were taken from published experiments; no new functional repair assays were performed.

Acknowledgments

The author thanks the investigators, data curators and contributors to CPTAC, CCLE, DepMap and GDSC, and the researchers who made their functional nucleotide excision repair classifications available for secondary analysis.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Conflict of interest

The author declares no competing interests.

Author contributions

Achab-Awelana Ernest: Conceptualization, methodology, data curation, formal analysis, investigation, visualization, writing – original draft, and writing – review and editing.

Ethics and consent

This study performed secondary analysis of publicly available, non-identifiable molecular and cell-line research data and published functional classifications. No new human participants or animals were enrolled, and no identifiable individual-level information was collected. Consent for publication is not applicable.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During preparation of this manuscript, the author used OpenAI ChatGPT to assist with code drafting, exploratory computational analyses, literature identification, figure preparation, manuscript organization and language editing. Outputs were reviewed and edited, and the author takes full responsibility for the accuracy, originality and integrity of the submitted work.

References

1. Compe E, Egly JM (2012) TFIIH: when transcription met DNA repair. Nat Rev Mol Cell Biol 13:343–354. 10.1038/nrm3350
2. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JHJ (2014) Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol 15:465–481. 10.1038/nrm3822
3. Taggart JC, Zauber H, Selbach M, Li GW, McShane E (2020) Keeping the proportions of protein complex components in check. Cell Syst 10:125–132. 10.1016/j.cels.2020.01.004
4. Gonçalves E et al (2017) Widespread post-transcriptional attenuation of genomic copy-number variation in cancer. Cell Syst 5:386–398e4. 10.1016/j.cels.2017.08.013
5. Zhang B et al (2014) Proteogenomic characterization of human colon and rectal cancer. Nature 513:382–387. 10.1038/nature13438
6. Dou Y et al (2020) Proteogenomic characterization of endometrial carcinoma. Cell 180:729–748e26. 10.1016/j.cell.2020.01.026
7. Giglia-Mari G et al (2004) A new, tenth subunit of TFIIH is responsible for the DNA repair syndrome trichothiodystrophy group A. Nat Genet 36:714–719. 10.1038/ng1387
8. Olaussen KA et al (2006) DNA repair by ERCC1 in non-small-cell lung cancer and cisplatin-based adjuvant chemotherapy. N Engl J Med 355:983–991. 10.1056/NEJMoa060570
9. Friboulet L et al (2013) ERCC1 isoform expression and DNA repair in non–small-cell lung cancer. N Engl J Med 368:1101–1110. 10.1056/NEJMoa1214271
10. Stojkova O et al (2026) Identification of an ERCC2 mutation-associated mutational signature of nucleotide excision repair deficiency in targeted panel sequencing data. JCO Precis Oncol 10:e2600329. 10.1200/PO-26-00329
11. Lindgren CM et al (2021) Simplified and unified access to cancer proteogenomic data. J Proteome Res 20:1902–1910. 10.1021/acs.jproteome.0c00919

12. Nusinow DP et al (2020) Quantitative proteomics of the Cancer Cell Line Encyclopedia. Cell 180:387–402e16. 10.1016/j.cell.2019.12.023
13. Tsherniak A et al (2017) Defining a Cancer Dependency Map. Cell 170:564–576e16. 10.1016/j.cell.2017.06.010
14. Iorio F et al (2016) A landscape of pharmacogenomic interactions in cancer. Cell 166:740–754. 10.1016/j.cell.2016.06.017
15. Jabeen A et al (2026) SLFN11 enhances cisplatin sensitivity in pediatric cancer via activation of stress-response and suppression of survival pathways. Br J Cancer. 10.1038/s41416-026-03571-2
16. COSMIC (2026) SBS8 mutational signature database, v3.6 https://cancer.sanger.ac.uk/signatures/sbs/sbs8/(accessed 8 November 2026).
17. Ribeiro-Silva C, Aydin ÖZ, Mesquita-Ribeiro R et al (2018) DNA damage sensitivity of SWI/SNF-deficient cells depends on TFIIH subunit p62/GTF2H1. Nat Commun 9:4067. 10.1038/s41467-018-06402-y
18. Rajkumar-Calkins AS, Szalat R, Dreze M et al (2019) Functional profiling of nucleotide excision repair in breast cancer. DNA Repair (Amst) 82:102697. 10.1016/j.dnarep.2019.102697
19. Wei R, Dai H, Zhang J et al (2021) A gene expression signature to predict nucleotide excision repair defects and novel therapeutic approaches. Int J Mol Sci 22:5008. 10.3390/ijms22095008
20. Vaughn CM, Selby CP, Yang Y, Hsu DS, Sancar A (2020) Genome-wide single-nucleotide resolution of oxaliplatin–DNA adduct repair in drug-sensitive and -resistant colorectal cancer cell lines. J Biol Chem 295:7584–7596. 10.1074/jbc.RA120.013347

Figures

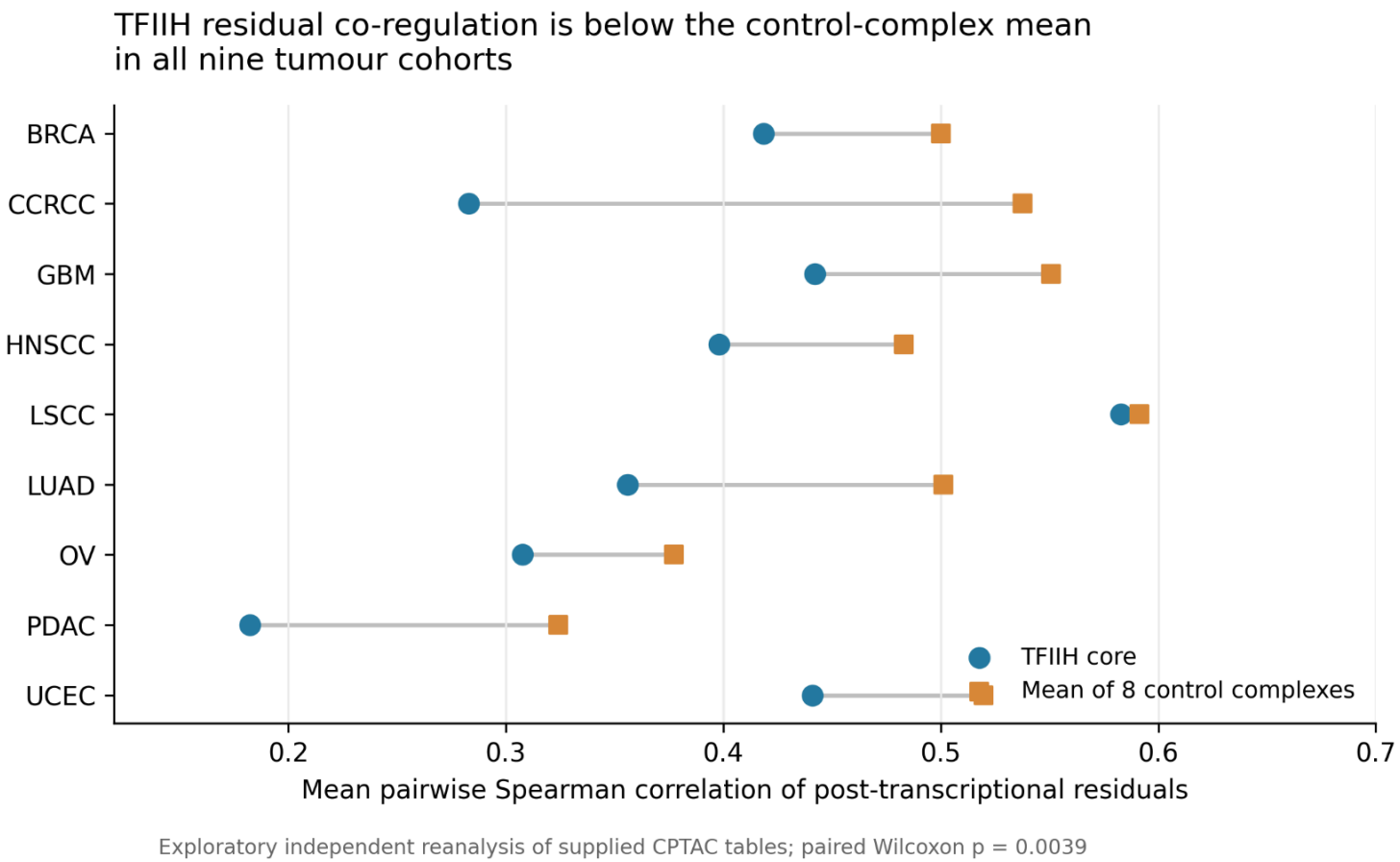


Figure 1

Independent specificity benchmark. Each segment joins the TFIIH core residual co-regulation value with the mean of eight non-TFIIH complex values within the same CPTAC cohort. Correlations are descriptive, and the paired cohort test is an exploratory reanalysis.

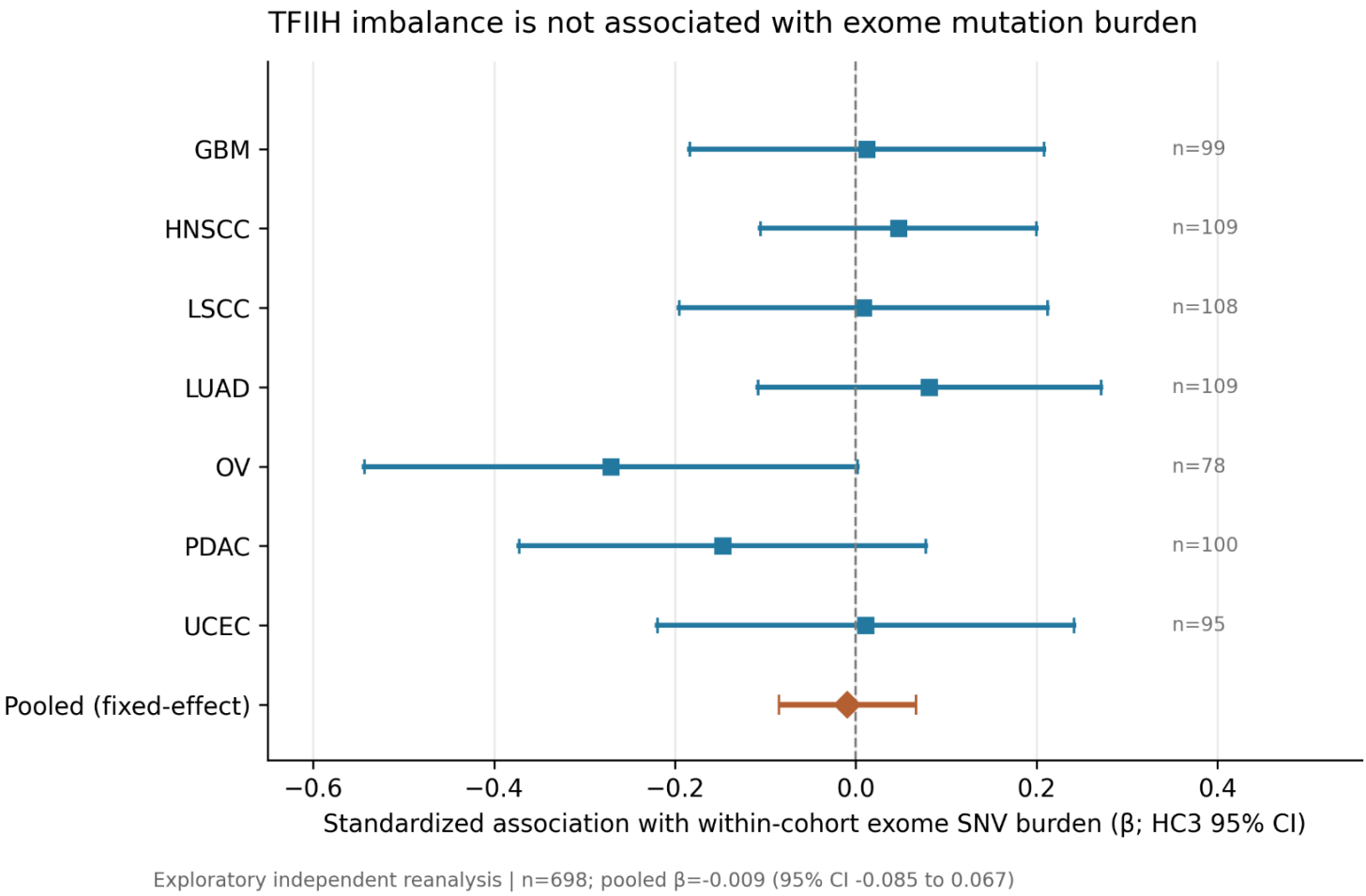
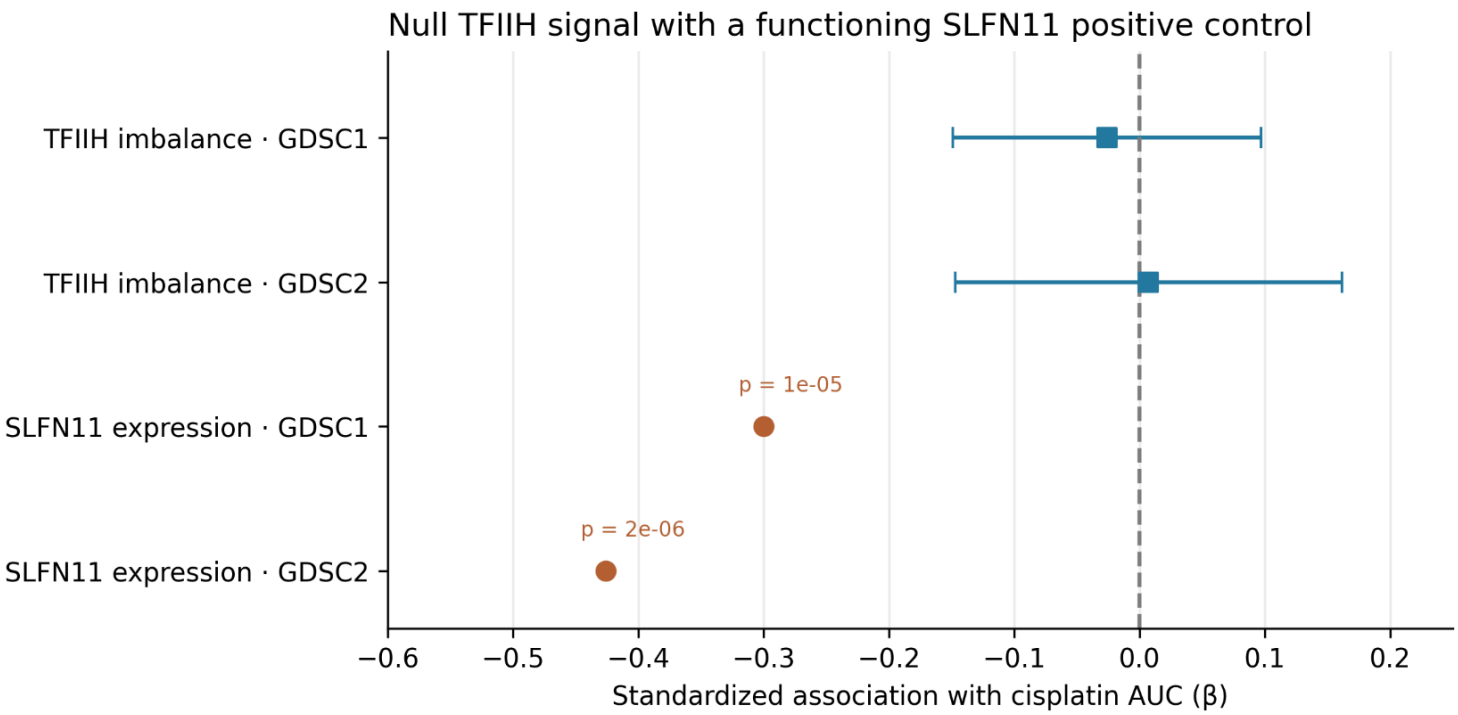


Figure 2

Independent tumour burden associations. Each cohort estimate is the HC3-robust standardized TFIIH imbalance coefficient for rank-normalized exome SNV counts after adjustment for purity, proliferation and mean TFIIH RNA. Pooled result uses an inverse-variance fixed-effect model; 95% intervals are approximate.



TFIIH: HC3-robust 95% CIs; SLFN11: positive-control coefficients (CIs not plotted)

Figure 3

TFIIH imbalance versus SLFN11 cisplatin association. TFIIH coefficients include HC3-robust 95% CIs; SLFN11 markers display the coefficients of the parallel positive-control models (confidence intervals omitted from this figure). Values are standardized within each GDSC dataset.

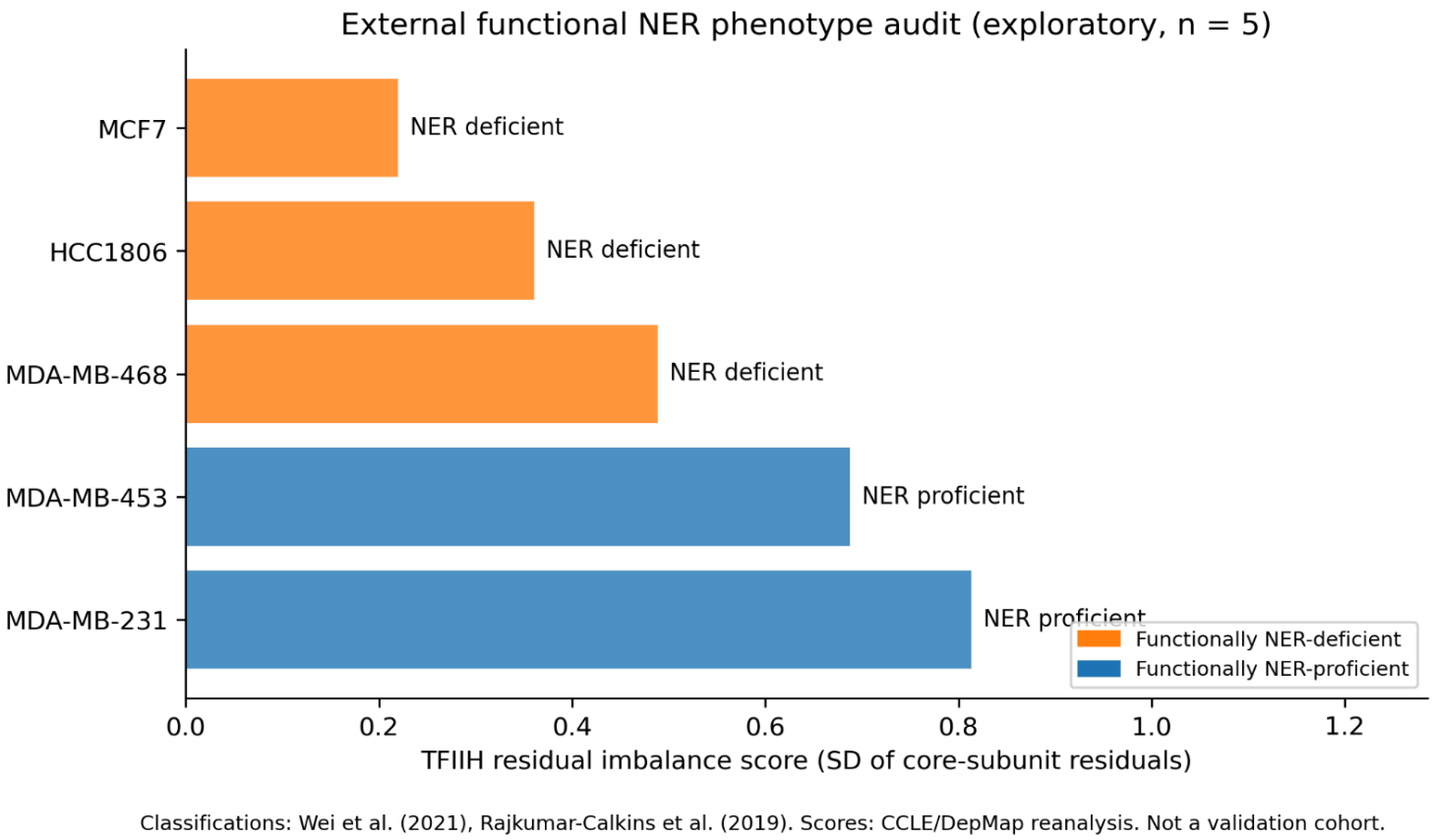


Figure 4

Exploratory published functional NER phenotype triangulation. TFIIH residual imbalance of five CCLE/DepMap models matched to experimental NER classifications in two independent papers. Missing scores in SKBR3 and MDA-MB-361 were not imputed. The observed reverse ranking is not a validated inverse biomarker effect because n = 5 and phenotype assays were not harmonized.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DNARepairSupplementaryInformation.docx