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.

Cross-gene kinase domain hotspot analysis with AI pathogenicity prediction identifies FGFR2 D650G as a candidate actionable variant

Teruki Tsunashima
National Cancer Centre Japan
Kosuke Maruyama
National Cancer Centre Japan
Emi Ogawa
National Cancer Centre Japan
Nguyen Thai Le
National Cancer Centre Japan
Takashi Kohno
National Cancer Centre Japan
Takashi Nakaoku
tnakaoku@ncc.go.jp

National Cancer Centre Japan

Article

Keywords:

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10221428/v1

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

Additional Declarations: No competing interests reported.

Abstract

Variants of unknown significance (VUSs) constitute most variants detected by cancer genomic profiling, limiting the clinical application of precision oncology. We developed a systematic framework that integrates cross-gene mutation mapping across 41 receptor tyrosine kinase genes with AI-driven pathogenicity prediction. By mapping 32,462 kinase-domain mutations from the AACR Project GENIE database onto a consensus sequence, we identified conserved cross-gene hotspots. AI-based stratification using AlphaMissense and PrimateAI-3D identified the activation-loop hotspot as a reservoir of uncharacterized candidate drivers, in contrast to the passenger-enriched back-loop hotspot. We prioritized and functionally validated FGFR2 D650G, a VUS that exhibited constitutive kinase activity and sensitivity to the FDA-approved FGFR inhibitors erdafitinib and futibatinib comparable to those of the established driver FGFR2 N549K. This computation-to-validation approach may facilitate the prioritization and functional assessment of overlooked VUSs as candidate actionable therapeutic targets, highlighting the value of integrating evolutionary conservation with AI prediction.

Introduction

Next-generation sequencing has transformed cancer care by enabling the identification of oncogenic driver mutations that are amenable to targeted molecular therapies[1–3]. This precision oncology paradigm relies on accurate functional annotation of genetic variants to guide treatment decisions. However, a critical bottleneck remains: approximately 90% of variants detected through clinical sequencing are classified as variants of unknown significance (VUSs), creating a substantial barrier to translating genomic data into effective targeted therapies[4–6]. Therefore, distinguishing the relatively small number of functionally relevant driver mutations from the vast majority of neutral passenger mutations remains a central challenge in precision oncology.

Traditionally, driver mutations have been identified based on their recurrent occurrence within individual genes across cancer cohorts. However, this gene-centric perspective may overlook important dimensions of oncogenesis. As proposed by Miller et al., selective pressures in cancer may operate not only at the level of individual genes but also on specific molecular functions conserved across gene families[7]. They demonstrated that cancer-associated mutations frequently cluster within conserved protein domains, particularly among receptor tyrosine kinases (RTKs), which are key regulators of cell proliferation, differentiation, and survival. The kinase domain of RTKs, which mediates ATP-dependent phosphorylation and downstream signal transduction, is a well-established hotspot for activating mutations in genes such as *EGFR*, *KIT*, *FLT3*, *RET*, and *FGFR2*; these findings have directly supported the development of targeted inhibitors[8–12]. These observations suggest that mutations occurring at functionally equivalent positions across distinct RTK genes may converge on shared oncogenic mechanisms.

A cross-gene analytical approach is therefore needed to systematically uncover such patterns. Recent advances in artificial intelligence have also provided tools for predicting the pathogenicity of genetic

variants. AI-based pathogenicity prediction methods, including AlphaMissense, which leverages deep learning informed by protein structure and evolutionary conservation, and PrimateAI-3D, which is trained on genetic variation across 233 primate species, have shown utility for variant classification without relying on potentially biased clinical annotations[13,14]. Consistent with this approach, previous work has used real-world data and AlphaMissense to improve the interpretation of VUSs[15,16].

In this study, we hypothesized that integrating cross-gene hotspot analysis with structure-informed AI-based pathogenicity prediction would enable effective filtering of the extensive VUS landscape to identify candidate functionally relevant driver mutations. We developed a framework that maps tens of thousands of RTK mutations from the AACR Project GENIE cohort onto a consensus kinase-domain sequence, identifies cross-gene mutation hotspots, and prioritizes candidate VUSs using AlphaMissense and PrimateAI-3D. Through this systematic approach, we identified and functionally validated *FGFR2* D650G as a previously unrecognized candidate driver mutation with sensitivity to FGFR inhibitors, thereby demonstrating the potential of our strategy to expand the repertoire of candidate actionable targets in precision oncology.

Results

Pan-cancer mapping of kinase-domain mutations identifies conserved hotspots

To systematically identify functionally constrained regions across the receptor tyrosine kinase (RTK) family, we analyzed 32,462 missense mutations and in-frame insertions/deletions (indels) across 41 RTK genes using data from the AACR Project GENIE database versions 14.0-public and 14.1-public[17,18]. By mapping these variants onto a consensus kinase-domain sequence derived from multiple sequence alignment, we generated a unified mutational landscape comprising 403 aligned residues (ARs) (Fig. 1A **and Supplementary Table S1**). The cross-gene analysis showed that regions with higher pathogenicity scores, as assessed using AlphaMissense (Fig. 1C **and Supplementary Table S2**) and PrimateAI-3D (Fig. 1D **and Supplementary Table S3**), tended to coincide with highly conserved amino-acid positions (Fig. 1B). Consistent with this finding, the observed somatic mutations were enriched at conserved sites, indicating that mutations preferentially occurred at positions with high pathogenicity scores rather than being randomly distributed across the kinase domain. After excluding non-reproducible solitary mutations, we identified six major mutation hotspots, each accumulating more than 500 mutations across multiple cancer types: AR 47–48 within the β-sheet III region, AR 78 in the back-loop, AR 99 at the gatekeeper residue, and AR 271 and AR 274 within the activation loop (Fig. 1A and 1E). The concordant alignment of well-established driver mutations, such as *EGFR* L858R at AR 271 and *EGFR* T790M at AR 99, supported the ability of this consensus framework to capture functionally equivalent positions across diverse kinases.

Characterization of hotspots reveals distinct driver/VUS compositions

We next examined the genetic composition of these hotspots to prioritize regions harboring potentially overlooked candidate driver mutations (Fig. 2A). We deprioritized three hotspots based on predefined criteria. AR 48 (β-sheet III) and AR 271 (activation loop) were each dominated by mutations originating from a single gene—EGFR exon 19 deletions and EGFR L858R, respectively—and therefore lacked cross-gene diversity. Although AR 99 (gatekeeper) harbored mutations from multiple RTK genes, it was predominantly composed of EGFR T790M and analogous gatekeeper-type substitutions, which represent a well-characterized resistance mechanism that is functionally distinct from gain-of-function driver mutations. These three hotspots therefore offered limited opportunity for novel driver discovery[19–21]. We therefore further examined the composition of the remaining hotspots—AR 47, AR 78, and AR 274—each of which harbored mutations from multiple RTK genes (Fig. 2B). Although AR 47 (β-sheet III) harbored mutations from diverse RTK genes, its composition was predominantly accounted for by EGFR exon 19 deletions (E746_A750 del) in non-small-cell lung cancer (Fig. 2B **and Supplementary Figure S1A**). By contrast, AR 78 (back-loop) and AR 274 (activation loop) displayed heterogeneous compositions, with contributions from more than 15 distinct RTK genes and a mixture of established driver mutations and VUSs spanning a broad range of cancer types (Fig. 2B **and Supplementary Figure S1B, C**). Specifically, AR 78 harbored known drivers, such as FGFR1 N546K, alongside recurrent VUSs, including *ALK* R1181C[22,23]. Similarly, AR 274 contained established drivers, such as KIT D816V and EGFR L861Q, that co-occurred with numerous VUSs, including *FGFR4* G636S, *FGFR2* D650G, and *PDGFRB* D850V[24,25]. This heterogeneous composition suggests that these two hotspots represent conserved, potentially functionally relevant positions that are recurrently targeted by mutations across a broad spectrum of kinases.

AI-based prediction distinguishes candidate functional drivers from passengers

To distinguish candidate functional driver mutations from passenger VUSs within these prioritized hotspots, we applied the pathogenicity prediction algorithms AlphaMissense and PrimateAI-3D[13,14]. This analysis revealed a clear predicted functional divergence between AR 78 (Fig. 3A) and AR 274 (Fig. 3B). Evaluation using both tools classified annotated driver mutations as pathogenic. At the back-loop hotspot (AR 78), despite the high mutation counts, most VUSs (e.g., ALK R1181C, EPHB1 R682C, and NTRK1 R574C) were categorized as likely benign by AlphaMissense (e.g., AlphaMissense score < 0.34), suggesting that this region may tolerate passenger mutations (Fig. 3A, **Supplementary Figure S2A and S2B, and Supplementary Tables S2 and S3**). In contrast, at the activation-loop hotspot (AR 274), 85% of VUSs were classified as likely pathogenic (AlphaMissense score > 0.564). VUSs such as FGFR2 D650G (score 0.89) and PDGFRB D850V (score 0.86) received high pathogenicity scores comparable to those of established driver mutations (Fig. 3B, **Supplementary Figure S2C and S2D, and Supplementary Tables**

S2 and S3). This AI-based stratification supported deprioritization of AR 78, which contained variants predicted to have limited functional impact, and prioritization of AR 274 as a reservoir of previously uncharacterized candidate driver mutations, such as *FGFR2* D650G and *PDGFRB* D850V.

Functional and therapeutic assessment of FGFR2 D650G

We next functionally assessed the oncogenic potential and therapeutic sensitivity of the prioritized candidates *FGFR2* D650G and *PDGFRB* D850V. In vitro functional assays using HEK293H cells demonstrated that both *FGFR2* D650G and *PDGFRB* D850V conferred ligand-independent autophosphorylation and downstream ERK activation at levels comparable to those of the established driver mutations FGFR2 N549K and PDGFRB N666K relative to wild-type (WT) controls (Fig. 4A and 4B). Furthermore, analysis of clinical sequencing data from the GENIE dataset revealed that FGFR2 D650G exhibited a significantly elevated variant allele frequency (mean VAF: 0.438, P = 0.0086), consistent with positive selection during tumor evolution (**Supplementary Figure S3 and Supplementary Table S5**)[26]. FGFR2 D650G phosphorylation was suppressed by treatment with the FDA-approved FGFR inhibitors futibatinib and erdafitinib, with IC_{50} values of approximately 45 nM for futibatinib (vs. 122 nM for N549K) and approximately 88 nM for erdafitinib (vs. 1,603 nM for N549K) (Fig. 4C)[12,27]. Because our analysis focused on mutation accumulation within the kinase domain, we performed functional experiments using recombinant FGFR2 kinase-domain proteins to assess the efficacy of *FGFR* kinase inhibitors. These experiments showed that inhibitor treatment suppressed phosphorylation of the recombinant *FGFR2* kinase domain harboring D650G as well as that harboring N549K (Fig. 4D). Taken together, these findings indicate that the FGFR2 D650G mutation exhibited ligand-independent activation and may represent a potential target for FGFR tyrosine kinase inhibitors.

Discussion

In this study, we addressed a central challenge in precision oncology—the interpretation of VUSs—by establishing a systematic framework that integrates cross-gene mutation mapping with AI-driven pathogenicity prediction. Whereas traditional gene-centric analyses may fail to detect rare driver mutations because of limited statistical power and sample availability, our approach leverages functional redundancy within the kinase domain to aggregate statistical power across the RTK family. By redefining somatic mutations according to their homologous positions within a consensus kinase sequence, referred to as aligned residues, rather than according to gene-specific coordinates, we identified candidate driver mutations that might otherwise be dismissed as background noise. The identification and functional validation of *FGFR2* D650G as a candidate drug-sensitive driver mutation provide proof of concept that this multimodal strategy can help expand the landscape of candidate actionable therapeutic targets in precision oncology.

Our findings underscore the value of integrating hotspot analysis with AI-based pathogenicity prediction. Previous studies, including those by Miller *et al.*, have highlighted the importance of domain-level

mutation hotspots in cancer[7]. However, our analysis suggests that physical clustering alone may be insufficient to define pathogenicity. For example, the back-loop hotspot (AR 78) accumulated a high number of mutations, many of which were predicted to be benign passengers. By overlaying AlphaMissense and PrimateAI-3D predictions onto these cross-gene hotspots, we characterized their internal composition and distinguished predicted low-functional-impact hotspots, such as AR 78, from predicted high-functional-impact hotspots, such as AR 274. This stratification may help prioritize candidates for experimental validation and highlights the utility of AI-based approaches as high-throughput filters for identifying variants less likely to have functional effects.

The identification of *FGFR2* D650G may have clinical implications after further validation. Our functional assays demonstrated that this mutant was constitutively active and sensitive to the FDA-approved FGFR inhibitors erdafitinib and futibatinib, similar to the well-characterized *FGFR2* N549K driver[12,27]. This inhibitor sensitivity raises the possibility that *FGFR2* D650G, currently classified as a VUS, could warrant further evaluation as a candidate biomarker for *FGFR*-targeted therapies. Collectively, these findings highlight the capacity of our workflow to identify overlooked variants that may represent promising candidates for further precision-oncology investigation.

Although this study focused on the RTK family, the principles underlying our framework may be applicable to other enzyme families with conserved catalytic domains, including serine/threonine kinases, phosphatases, and additional signaling molecules. One limitation of the present study is its reliance on computational predictions, which were supported by the functional validation of selected candidates but still require broader experimental and clinical confirmation for translation. In addition, because our experimental validation was restricted primarily to *FGFR2* D650G, integration with higher-throughput functional assays may enable the systematic identification and characterization of a broader spectrum of candidate driver mutations. In conclusion, reanalyzing existing genomic datasets using AI-based strategies may help prioritize rare genetic alterations for functional validation and therapeutic investigation in precision oncology.

Materials and methods

Data acquisition and mutation selection

Mutation data were obtained from the American Association for Cancer Research (AACR) Project GENIE, a publicly available pan-cancer registry containing clinical-grade genomic data from more than 100,000 patients with cancer across more than 100 tumor types[17,18]. We downloaded variant information for 58 genes belonging to the HUGO Gene Nomenclature Committee (HGNC) receptor tyrosine kinase (RTK) group and RTK subgroup from GENIE cohort versions 14.0-public and 14.1-public accessed in October 2024[28]. From these data, we extracted missense mutations and in-frame insertions/deletions (indels) located within the annotated kinase domain of each gene. Kinase-domain boundaries were defined according to UniProt annotations[29]. After applying quality filters, including removal of duplicates and

variants with missing positional information, we retained 32,462 mutations across 41 RTK genes for subsequent analysis.

Multiple sequence alignment and consensus-sequence generation

Amino-acid sequences corresponding to the kinase domains of all 41 RTK genes were retrieved from UniProt. Multiple sequence alignment (MSA) was performed using the ClustalW algorithm (version 2.1) with default parameters[30,31]. The resulting alignment generated a consensus sequence of 403 positions, designated as aligned residues (ARs) numbered 1 through 403. Conserved structural motifs of the kinase domain, including the glycine-rich loop (P-loop), catalytic loop (HRD motif), DFG motif, and activation loop, were mapped onto the consensus sequence according to established kinase structural annotations[32].

Cross-gene mutation mapping and hotspot identification

Individual mutations were mapped to their corresponding AR positions based on the MSA. To prevent overestimation of clustering due to sequencing artifacts or germline polymorphisms, we included only mutations observed in two or more independent samples in aggregation analyses. Mutation hotspots were identified using two criteria: (i) the total mutation count at each AR position and (ii) the number of distinct RTK genes contributing mutations at that position. This two-dimensional evaluation was visualized using scatter plots, and positions accumulating > 500 total mutations from ≥ 15 genes were designated as significant cross-gene hotspots. For each identified hotspot, constituent mutations were annotated according to OncoKB classification as established driver mutations or variants of unknown significance (VUSs)[5].

AI-based pathogenicity prediction

Pathogenicity predictions for all possible single-amino-acid substitutions (19 missense variants × number of residues) at constituent gene positions within the identified hotspots were obtained using AlphaMissense[13]. AlphaMissense scores were downloaded from the publicly available AlphaMissense database (https://alphamissense.hegelab.org). Variants were classified according to the published thresholds as likely pathogenic (score ≥ 0.564), ambiguous (score 0.34–0.564), or likely benign (score ≤ 0.34). For comparison, we also obtained predictions from PrimateAI-3D[14], a semi-supervised three-dimensional convolutional neural network trained on 4.3 million common primate variants.

Variant allele frequency analysis

Variant allele frequencies (VAFs) for mutations at significant hotspots were extracted from the GENIE dataset. VAF distributions are shown as box plots. Mean VAFs were compared between established driver mutations and VUSs using two-tailed Student's t-tests. Statistical significance was defined as $P < 0.05$. Only VUSs with ≥ 10 independent observations were included in statistical comparisons.

Plasmid construction and mutagenesis

The wild-type *FGFR2* expression plasmid pHAGE-FGFR2 (Addgene plasmid #116741) was obtained from Addgene[33]. Site-directed mutagenesis was performed using the QuikChange Lightning Site-Directed Mutagenesis Kit (Agilent Technologies, #210518) according to the manufacturer's protocol. The D650G mutation (GAT→GGT) was introduced to generate pHAGE-FGFR2-D650G, and the N549K mutation (AAC→AAG) was introduced as a positive control. The kinase-dead K517R mutation was generated as a negative control[34]. All constructs were verified by Sanger sequencing. Plasmid DNA was prepared using the HiSpeed Plasmid Maxi/Midi Kit (QIAGEN, #12643). The primer sequences are provided in Supplementary Table S6.

Cell culture and transient transfection

HEK293H cells (Thermo Fisher Scientific, #11631017) were maintained in Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 U/mL penicillin, and 100 µg/mL streptomycin at 37°C in a humidified atmosphere containing 5% CO_2. For transfection experiments, cells were seeded in 6-well plates at 0.5 × 10^6 cells/well in DMEM containing 5% FBS and cultured overnight to achieve 70%–90% confluence. Transfections were performed using Lipofectamine 3000 Reagent (Thermo Fisher Scientific, #3000015) according to the manufacturer's protocol. Plasmid DNA (1.0 µg for wild-type FGFR2 and 2.0 µg for mutant constructs) and P3000 Reagent (2 µL per µg DNA) were diluted in 125 µL Opti-MEM. Lipofectamine 3000 (3.75 µL) was diluted in 125 µL Opti-MEM. The solutions were combined, incubated for 10 min at room temperature, and added to cells in serum-free DMEM. At 24 h post-transfection, the medium was replaced with serum-free DMEM. Cells were harvested 48 h post-transfection.

Protein production

Recombinant baculovirus was generated using the flashBAC ULTRA system (Oxford Expression Technologies, Oxford, UK). Briefly, Sf9 insect cells (Thermo Fisher Scientific, #12659017) in Sf-900 II serum-free medium (Thermo Fisher Scientific) were seeded at a density of 1 × 10^6 cells/well in 6-well plates. The cells were co-transfected with the pBacPAK-GST-FGFR2 transfer plasmid encoding residues 456–768 and flashBAC ULTRA viral DNA using Cellfectin II transfection reagent (Thermo Fisher Scientific) in Sf-900 II serum-free medium. The resulting P1 viral supernatant was harvested 5 days post-transfection. To generate a high-titer stock, the virus was amplified for two additional passages (P2 and P3) by infecting larger-scale suspension cultures of Sf9 cells. The final P3 viral stock was harvested by centrifuging the culture at 12,000 × g for 10 min at 4°C, and the supernatant was stored at 4°C protected from light.

For protein production, Sf9 cells in suspension culture were infected with the P3 baculovirus stock. Cells were harvested 72 h post-infection and lysed by sonication. The GST-tagged FGFR2 kinase domain (KD) was purified from the clarified lysate by incubation with glutathione Sepharose 4B resin (Cytiva,

Marlborough, MA, USA). The purified proteins were stored at 4°C and used for biochemical assays within 72 h.

Drug treatment

For inhibitor-sensitivity assays, the FDA-approved FGFR inhibitors erdafitinib (#S8401) and futibatinib (#S8848) were purchased from Selleck Chemicals and dissolved in DMSO to prepare 10 mM stock solutions. At 24 h post-transfection, the culture medium was replaced with serum-free DMEM containing inhibitors at the indicated concentrations (0, 10, 100, 1,000, and 10,000 nM). The final DMSO concentration was maintained at 0.1% across all conditions. Cells were harvested after 24 h of drug exposure.

Western blot analysis

Cells were washed twice with ice-cold PBS and lysed in RIPA buffer supplemented with Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher Scientific, #78442). Protein concentrations were determined using the Pierce 660 nm Protein Assay Reagent (Thermo Fisher Scientific, #22660). Equal amounts of protein (20 μg) were resolved on 4%–20% Mini-PROTEAN TGX Precast Protein Gels (Bio-Rad, #4561086) and transferred to PVDF membranes (Bio-Rad, #1704157). Membranes were blocked with PVDF Blocking Reagent (TOYOBO, #NYPBR01) for 1 h at room temperature. Primary antibodies, all from Cell Signaling Technology and diluted 1:1,000, were as follows: FGFR2 (#11835), phospho-FGFR (Tyr653/654; #3471), p44/42 MAPK (Erk1/2; #9102), phospho-p44/42 MAPK (Thr202/Tyr204; #4370), and β-actin (#3700). HRP-conjugated secondary antibodies (anti-rabbit IgG, #7074; anti-mouse IgG, #7076) were diluted 1:6,000. Protein bands were visualized using enhanced chemiluminescence reagent and imaged using the iBright Imaging System 1500 FL (Thermo Fisher Scientific).

Declarations

Data availability

The mutation data used in this study are publicly available through the AACR Project GENIE database (https://www.synapse.org/genie). AlphaMissense prediction scores are available at https://alphamissense.hegelab.org. PrimateAI-3D prediction scores are available at https://primateai3d.basespace.illumina.com. All other data supporting the findings of this study are available from the corresponding author upon reasonable request. The datasets are provided in Supplementary Tables.

Code availability

Custom scripts used for cross-gene mutation mapping and analysis are available from the corresponding author upon reasonable request.

Competing interests

The authors declare no financial or non-financial competing interests.

Author Contribution

T.T. performed the computational analyses and functional experiments and wrote the manuscript. K.M. contributed to the computational analyses. E.O. contributed to the acquisition of in vitro data. N.T.L. contributed to the acquisition of in vitro data. T.K. supervised the study and revised the manuscript. T.N. conceived and designed the study, supervised the research, and revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgement

We thank the AACR Project GENIE consortium and participating institutions for making genomic and clinical sequencing data available. This work was supported by the Japan Agency for Medical Research and Development (AMED; JP22ck0106721h), Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology of Japan (KAKENHI; JP24K22078), JST FOREST Program (Grant Number JPMJFR245D, Japan), the SGH Cancer Research Grant, and the Research Grant of the Princess Takamatsu Cancer Research Fund awarded to T.N.

References

1. Tsimberidou, A. M., Fountzilas, E., Nikanjam, M. & Kurzrock, R. Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treat. Rev.* 86, 102019 (2020). 10.1016/j.ctrv.2020.102019.
2. Kohno, T. et al. KIF5B-RET fusions in lung adenocarcinoma. *Nat. Med.* 18, 375–377 (2012). 10.1038/nm.2644.
3. Nukaga, S. et al. Mutation of CMTR2 in lung adenocarcinoma alters RNA alternative splicing and reveals therapeutic vulnerabilities. *Nat. Commun.* 16, 9754 (2025). 10.1038/s41467-025-64821-0.
4. Tabata, J. et al. Novel calcium-binding ablating mutations induce constitutive RET activity and drive tumorigenesis. *Cancer Res.* 82, 3751–3762 (2022). 10.1158/0008-5472.CAN-22-0834.
5. Suehnholz, S. P. et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. *Cancer Discov.* 14, 49–65 (2024). 10.1158/2159-8290.CD-23-0467.

6. Murciano-Goroff, Y. R., Suehnholz, S. P., Drilon, A., Chakravarty, D. & Precision. Oncology: 2023. *Cancer Discov.* 13, 2525–2531 (2023).
7. Miller, M. L. et al. Pan-cancer analysis of mutation hotspots in protein domains. *Cell Syst.* 1, 197–209 (2015). 10.1016/j.cels.2015.08.014.
8. Maemondo, M. et al. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N. Engl. J. Med.* 362, 2380–2388 (2010). 10.1056/NEJMoa0909530.
9. Blay, J.-Y., Kang, Y.-K., Nishida, T. & von Mehren, M. Gastrointestinal stromal tumours. *Nat. Rev. Dis. Prim.* 7, 22 (2021). 10.1038/s41572-021-00254-5.
10. Negotei, C. et al. A review of FLT3 kinase inhibitors in AML. *J. Clin. Med.* 12, 6429 (2023). 10.3390/jcm12206429.
11. Kohno, T. et al. *RET* fusion gene: Translation to personalized lung cancer therapy. *Cancer Sci.* 104, 1396–1400 (2013). 10.1111/cas.12275.
12. Goyal, L. et al. Futibatinib for *FGFR2* -Rearranged intrahepatic cholangiocarcinoma. *N. Engl. J. Med.* 388, 228–239 (2023). 10.1056/NEJMoa2206834.
13. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* 381, eadg7492 (2023). 10.1126/science.adg7492.
14. Gao, H. et al. The landscape of tolerated genetic variation in humans and primates. *Science* 380, eabn8153 (2023). 10.1126/science.abn8197.
15. Kohno, T. et al. C-CAT: The national datacenter for cancer genomic medicine in Japan. *Cancer Discov.* 12, 2509–2515 (2022). 10.1158/2159-8290.CD-22-0417.
16. Yazaki, S. et al. AlphaMissense for identifying pathogenic missense mutations in DNA damage repair genes in cancer. *JCO Precis. Oncol.* 9, e2400908 (2025). 10.1200/PO-24-00908.
17. André, F. et al. AACR project GENIE: Powering precision medicine through an international consortium. *Cancer Discov.* 7, 818–831 (2017). 10.1158/2159-8290.CD-17-0151.
18. Pugh, T. J. et al. AACR project GENIE: 100,000 cases and beyond. *Cancer Discov.* 12, 2044–2057 (2022). 10.1158/2159-8290.CD-21-1547.
19. Batra, U., Biswas, B., Prabhash, K. & Krishna, M. V. Differential clinicopathological features, treatments and outcomes in patients with Exon 19 deletion and Exon 21 L858R EGFR mutation-positive adenocarcinoma non-small-cell lung cancer. *BMJ Open Respir. Res.* 10, e001492 (2023).
20. Holohan, C., Van Schaeybroeck, S., Longley, D. B. & Johnston, P. G. Cancer drug resistance: An evolving paradigm. *Nat. Rev. Cancer* 13, 714–726 (2013). 10.1038/nrc3599.
21. Yun, C.-H. et al. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proc. Natl Acad. Sci. USA* 105, 2070–2075 (2008). 10.1073/pnas.0709662105.
22. Lew, E. D., Furdui, C. M., Anderson, K. S. & Schlessinger, J. The precise sequence of FGF receptor autophosphorylation is kinetically driven and is disrupted by oncogenic mutations. *Sci. Signal.* 2 (2009). 10.1126/scisignal.2000021.

23. Engelhardt, S. et al. Frequent FGFR1 hotspot alterations in driver-unknown low-grade glioma and mixed neuronal-glial tumors. *J. Cancer Res. Clin. Oncol.* 148, 857–866 (2022). 10.1007/s00432-021-03906-x.
24. Arock, M. et al. KIT mutation analysis in mast cell neoplasms: Recommendations of the European Competence Network on Mastocytosis. *Leukemia* 29, 1223–1232 (2015). 10.1038/leu.2015.24.
25. Huang, S.-F. et al. High frequency of epidermal growth factor receptor mutations with complex patterns in non–small cell lung cancers related to gefitinib responsiveness in Taiwan. *Clin. Cancer Res.* 10, 8195–8203 (2004). 10.1158/1078 - 0432.CCR-04-1245.
26. Boscolo Bielo, L. et al. Variant allele frequency: A decision-making tool in precision oncology? *Trends Cancer* 9, 1058–1068 (2023). 10.1016/j.trecan.2023.08.011.
27. Loriot, Y. et al. Erdafitinib in locally advanced or metastatic urothelial carcinoma. *N. Engl. J. Med.* 381, 338–348 (2019). 10.1056/NEJMoa1817323.
28. Seal, R. L. et al. Genenames.org: The HGNC resources in 2023. *Nucleic Acids Res.* 51, D1003–D1009 (2023). 10.1093/nar/gkac888.
29. Bateman, A. et al. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* 53, D609–D617 (2025). 10.1093/nar/gkae1010.
30. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* 23, 2947–2948 (2007). 10.1093/bioinformatics/btm404.
31. Chenna, R. Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* 31, 3497–3500 (2003). 10.1093/nar/gkg500.
32. Kanev, G. K. et al. The landscape of atypical and eukaryotic protein kinases. *Trends Pharmacol. Sci.* 40, 818–832 (2019). 10.1016/j.tips.2019.09.002.
33. Ng, P. K.-S. et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* 33, 450–462.e10 (2018). 10.1016/j.ccell.2018.01.021.
34. Gartside, M. G. et al. Loss-of-function fibroblast growth factor Receptor-2 mutations in melanoma. *Mol. Cancer Res.* 7, 41–54 (2009). 10.1158/1541-7786.MCR-08-0021.

Figures

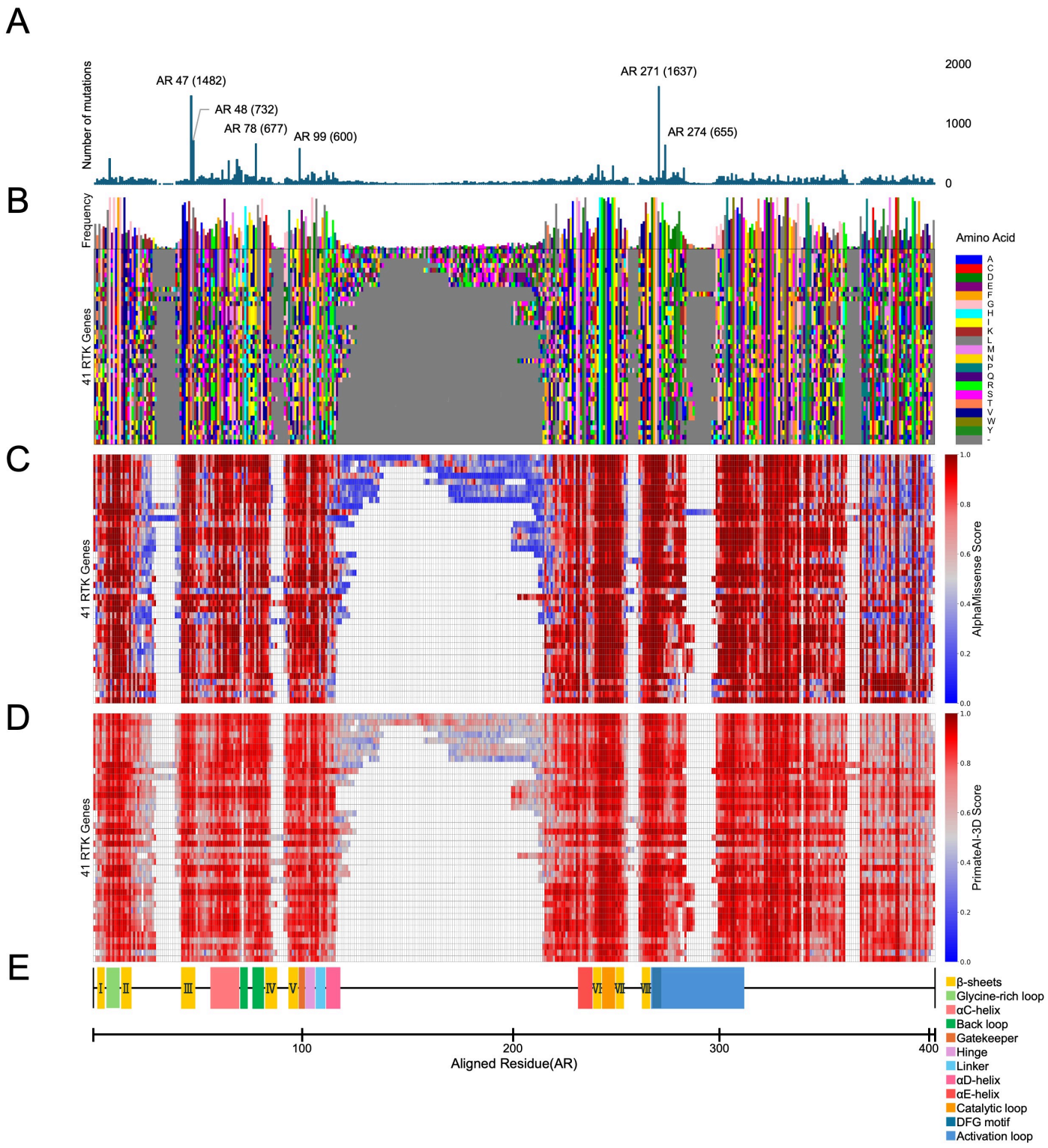


Figure 1

Kinase-domain consensus sequence and mutational landscape across 41 receptor tyrosine kinases (RTKs).

Consensus sequences were generated by multiple sequence alignment of kinase domains from 41 genes registered in the AACR Project GENIE database (v14.0-public and v14.1-public). (**A**) Distribution of

accumulated mutations across aligned residues (ARs) (n = 32,462); ARs with >500 mutations are indicated as hotspots. (**B**) Amino-acid composition at each AR in the 41-RTK consensus sequence (bottom), with the most frequent residue at each AR shown as a bar plot (top). (**C**) Heatmap of the mean AlphaMissense pathogenicity prediction scores for all 19 possible missense substitutions per position. (**D**) Heatmap of the mean PrimateAI-3D pathogenicity prediction scores for all possible single-nucleotide substitution patterns per position. (**E**) Common functional motifs in the kinase domain mapped to AR coordinates: β-sheets I (AR 5–7), II (AR 14–17), III (AR 43–48), IV (AR 79–82), V (AR 96–98), VI (AR 240–242), VII (AR 251–253), and VIII (AR 266); glycine-rich loop (AR 8–13); αC-helix (AR 59–69); back loop (AR 70–71 and 74–78); gatekeeper residue (AR 99); hinge region (AR 100–102); linker region (AR 103–106); αD-helix (AR 107–113); αE-helix (AR 235–239); catalytic loop (AR 243–250); DFG motif (AR 268–270); and activation loop (AR 271–311). AR positions 1–403 are shown along the bottom.

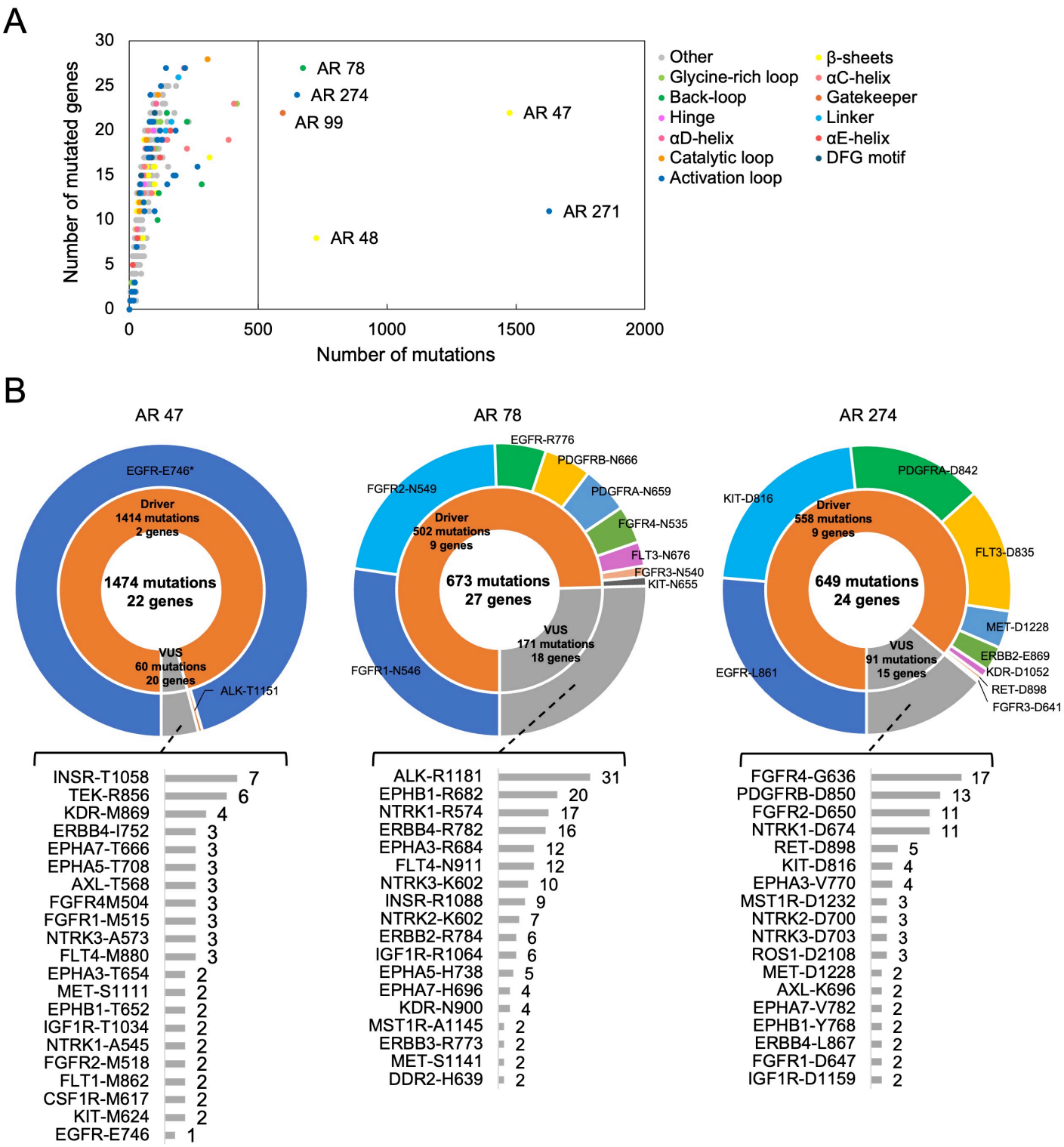


Figure 2

Hotspot residues mapped by mutation count and gene diversity.

(**A**) Scatter plot of hotspot residues showing the number of contributing genes and total mutation count. Mutation counts were aggregated across all cancer types; sites were included when at least two mutations mapped to the same AR. (**B**) OncoKB-based annotations of hotspot composition at AR 47

(left), AR 78 (middle), and AR 274 (right). The inner pie charts show the proportions of driver alterations and variants of unknown significance (VUSs), with the total numbers of mutations and genes indicated. The outer pie charts show the distribution of genes with mutated residues, and the bottom bar charts show genes with mutated residues classified as VUSs. To avoid overestimating aggregation and to ensure reproducibility, only variants detected in at least two samples were included in the analysis. The asterisk indicates E746_A750 del.

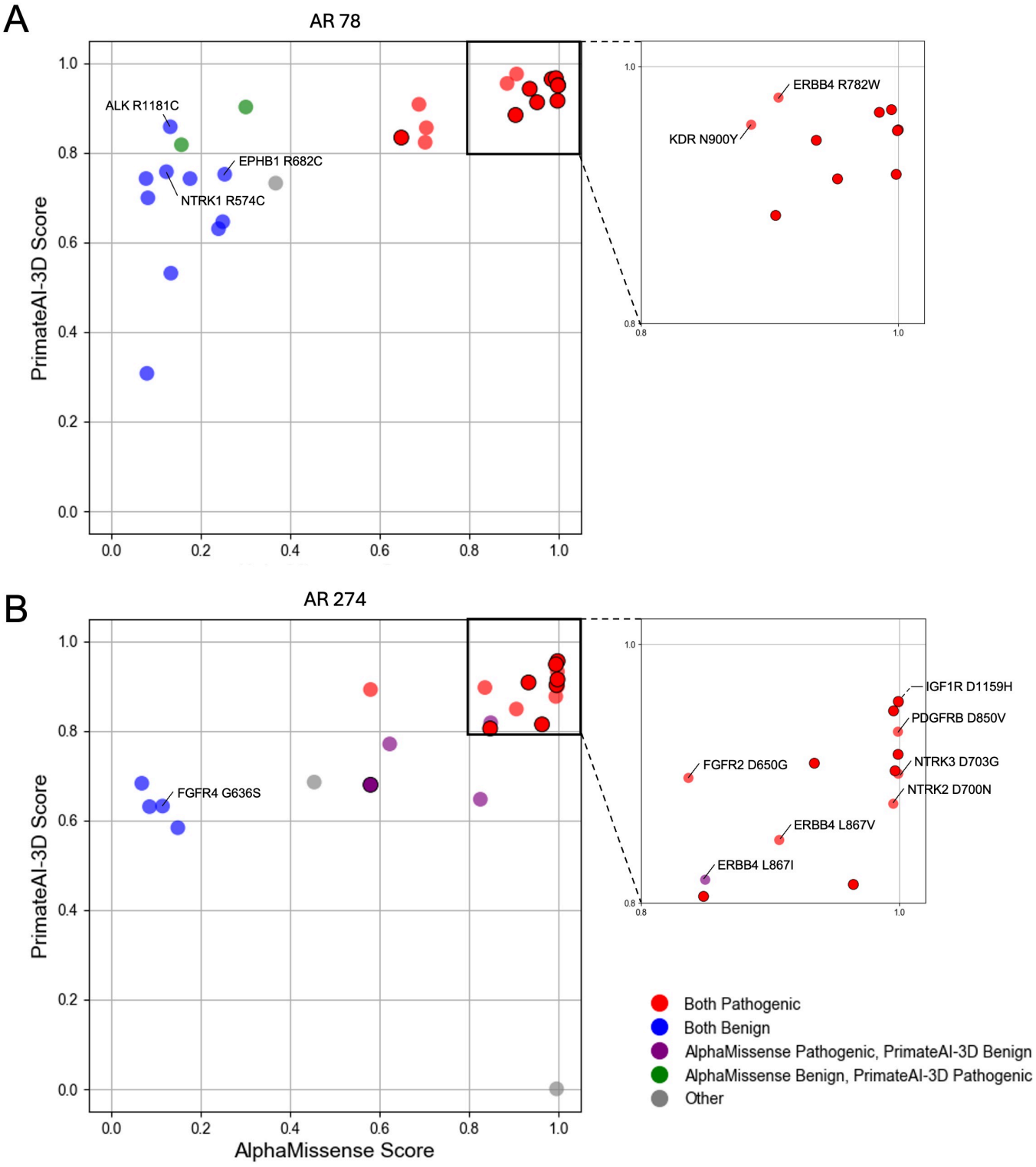


Figure 3

Distinguishing driver and passenger mutations using AI-driven pathogenicity predictions.
Scatter plots show the pathogenicity scores of mutations at AR 78 (**A**) and AR 274 (**B**) generated using AlphaMissense and PrimateAI-3D. Mutations are color-coded according to prediction outcomes: pathogenic by both tools (red), benign by both tools (blue), pathogenic by AlphaMissense and benign by PrimateAI-3D (purple), benign by AlphaMissense and pathogenic by PrimateAI-3D (green), and ambiguous or lacking predictions from either tool (gray). Driver mutations annotated by OncoKB are highlighted with black borders. VUSs with higher mutation counts and VUSs predicted as pathogenic by both tools are labeled.

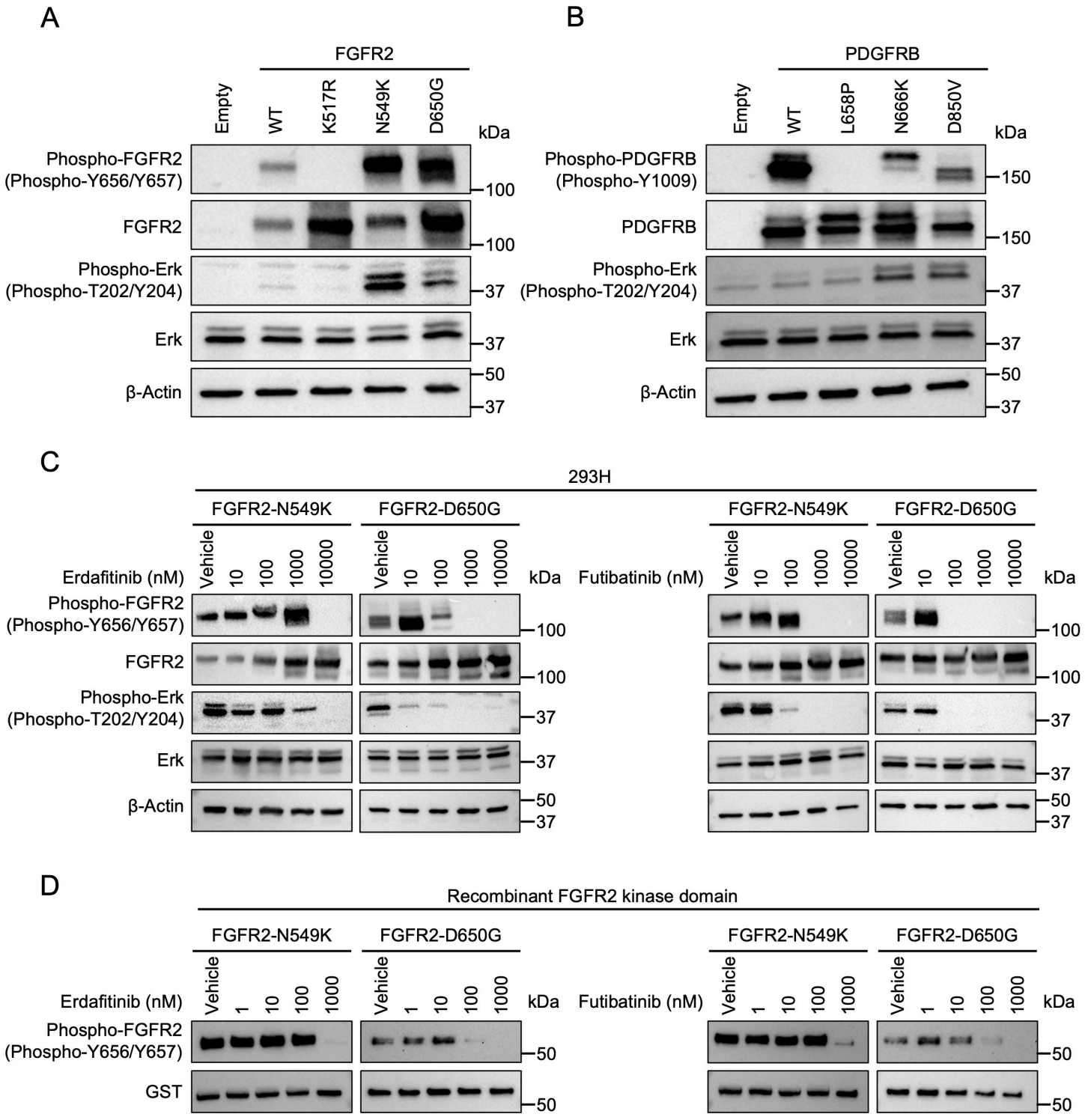


Figure 4

Gain-of-function activity and inhibitor sensitivity of FGFR2 D650G.

(**A**) Phosphorylation analysis of FGFR2 variants by immunoblotting. 293H cells were transfected with FGFR2 D650G, with wild-type (WT), K517R (loss-of-function), and N549K (gain-of-function) variants used as controls. FGFR2 phosphorylation and downstream ERK activation were assessed. (**B**) Phosphorylation analysis of PDGFRB variants by immunoblotting. 293H cells were transfected with PDGFRB D850V, with wild-type (WT), L658P (loss-of-function), and N666K (gain-of-function) variants

used as controls. PDGFRB phosphorylation and downstream ERK activation were assessed. (**C**) Effects of FGFR inhibitors on phosphorylation of activated FGFR2 variants. 293H cells expressing FGFR2 N549K or D650G were treated with erdafitinib or futibatinib, and inhibition of FGFR2 phosphorylation and ERK activation was evaluated by western blotting. (**D**) In vitro kinase-domain phosphorylation assay. Purified recombinant FGFR2 kinase-domain proteins (aa 456–768) harboring N549K or D650G, expressed using a baculovirus expression system, were treated with erdafitinib or futibatinib, and phosphorylation was assessed in a concentration-dependent manner.

Supplementary Files

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

- SupplementaryFigureS1.docx
- SupplementaryFigureS2.docx
- SupplementaryFigureS3.docx
- SupplementaryTableS1.xlsx
- SupplementaryTableS2.xlsx
- SupplementaryTableS3.xlsx
- SupplementaryTableS4.xlsx
- SupplementaryTableS5.xlsx
- SupplementaryTableS6.xlsx