Annals of human genetics

doi: 10.1111/ahg.12127

Twenty-One Novel EGFR Kinase Domain variants in Patients with Nonsmall Cell Lung Cancer

Marcia A. Hasenahuer[1], Gustavo Parisi[1], Marien Gautier[2], Alberto Lazarowski[2,3], Guillermo F. Bramuglia[2,3] and Maria Silvina Fornasari[1]*

[1]*Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Roque Saenz Peña 352, Bernal (B1876BXD), Buenos Aires, Argentina*

[2]*Fundación Investigar. Riobamba 429 Piso 15, Ciudad Autónoma de Buenos Aires (C1022AAT), Argentina*

[3]*Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 954, Ciudad Autónoma de Buenos Aires (C1113AAD), Argentina*

Summary

Somatic sequence variants in the epidermal growth factor receptor (EGFR) kinase domain are associated with sensitivity to tyrosine kinase inhibitors (TKIs) in patients with nonsmall cell lung cancer (NSCLC). Patients exhibiting sequence variants in this domain that produce kinase activity enhancement, are more likely to benefit from TKIs than patients with *EGFR* wild-type disease. Although most NSCLC *EGFR*-related alleles are concentrated in a few positions, established protocols recommend sequencing *EGFR* exons 18–21. In this study, 21 novel somatic variants belonging to such exons in adult Argentinean patients affected with NSCLC are reported. Of these, 18 were single amino acid substitutions (SASs), occurring alone or in combination with another genetic alteration (complex cases), one was a short deletion, one was a short deletion-short insertion combination, and one was a duplication. New variants and different combinations of previously reported variants were also found. Moreover, two of the reported SASs occurred in previously unreported positions of the EGFR kinase domain. In order to characterize the new sequence variants, physicochemical, sequence and conformational analyses were also performed. A better understanding of sequence variants in NSCLC may facilitate the most appropriate treatment choice for this complex disease.

Keywords: EGFR, tyrosine kinases, new sequence variations, lung cancer

Introduction

Gain-of-function or activating sequence variants of the epidermal growth factor receptor (*EGFR*) gene occur in some non-small cell lung cancers (NSCLC), leading to constitutive tyrosine kinase (TK) activity. Recognition of the relationship between these types of activating sequence variations and cancer has focused on an enormous quantity of research in molecular-targeted therapy (Zhang et al., 2010). Most of these disease associated sequence variants are found in exons 18–21 (Pao & Chmielecki, 2010) of the *EGFR* gene and the most frequent, occurring in approximately 90% of cases are short in-frame deletions in exon 19 and a single amino-acid substitution (SAS) L858R in exon 20 (Roengvoraphoj et al., 2013). Rare or uncommon sequence variations affect exons 18, 19, 20, and 21 (Beau-Faller et al., 2014). The frequency of NSCLC *EGFR*-associated sequence variants depends on the ethnic group studied (Arrieta et al., 2011; Bauml et al., 2013); moreover, position specific alleles are continuously reported in bibliographies and uploaded to different protein and cancer-related databases.

While several lines of evidence clearly show that patients harboring different kinds of *EGFR* activating sequence variants, and treated with tyrosine kinase inhibitors (TKIs), experience historically high survival rates, in other cases a nonresponse or insensitivity is observed (Sgambato et al., 2012). Consequently, it is important to establish the structural and functional connections between different variations and EGFR ligand interactions to understand differences in the expected response (Yasuda et al., 2013).

*Corresponding author: MARIA SILVINA FORNASARI, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes, Roque Saenz Peña 352, Bernal (B1876BXD), Buenos Aires, Argentina. Tel: ++54 11 43657100 ext 5657; Fax: ++54 11 43657132; E-mail: silvina333@gmail.com

Taking into account all these observations, we performed a large-scale sequencing study on the regional representation of *EGFR* variants in NSCLC Argentinean patients. Twenty-one new *EGFR* somatic variants as well as new combinations of previously reported variants, are presented and analyzed at protein level using sequence and structure-based methods. As a protein's native state is better represented by an ensemble of conformers, the effect of SASs should be evaluated on each of the available conformers of the protein. Following this approach, we found that the damaging effect of SASs on protein function evaluated in different conformations could help in the understanding of disease associated SASs (Juritz et al., 2012). For this reason, and to obtain a mechanistic analysis of the SAS effect, we explicitly studied the described SASs with regards to the conformational diversity of the EGFR protein.

Materials and Methods

Adult NSCLC affected patients (1865) were screened for somatic variant in exons 18–21 of the *EGFR* gene. This screening tends to determine treatment if the known response to a particular *EGFR* sequence variant is found. The tissue used for DNA extraction and analysis was obtained at the time of diagnosis. The included patients were characterized histologically as adenocarcinomas (NSCLC). The sequence of the *EGFR* gene corresponding to the kinase domain of the protein was amplified by polymerase chain reaction (PCR). Two separate PCRs, each with a corresponding pair of primers, were used to amplify exons 19 and 21 of the *EGFR* gene. The resulting PCR products were then subjected to direct sequencing using the same primers, and all variants were confirmed by sequences originating from both upstream and downstream primers. Of the patients analyzed, 210 exhibited nonsynonymous sequence variants in the sequenced exons that were checked in COSMIC (Forbes et al., 2011), Humsavar (Humsavar Database, Geneva, Switzerland: Swiss Institute of Bioinformatics, http://www.uniprot.org/docs/humsavar), ClinVar (ClinVar database, Release weekly,Bethesda, MD, USA: National Center for Biotechnology Information, US, National Library of Medicine, (http://www.ncbi.nlm.nih.gov/clinvar/), DMDM (Peterson et al., 2010) and bibliographic databases to verify previous reports. To evaluate whether these variants involved structural and/or functionally relevant positions, the structures of different conformers were extracted from the CoDNas database (Monzon et al., 2013). These conformers were the inactive and active EGFR kinase domain forms, as monomers or dimers (respectively, Protein Data Bank codes: 4i20, 1m14, 2gs7 or 3gt8, and 4g5p). Although the CoDNas database included all available conformers for EGFR, these were selected with regards to resolution, with the cocrystallized preferred over soaked; conformers with the lowest number of missing residues were also preferred. Relative Gibbs free energy differences for the unfolding, $\Delta G^{\text{unfolding}}$, between different SASs and wild type counterparts were estimated for all conformations using the FoldX program (Schymkowitz et al., 2005). Every time a missing residue was found in studied structures, its $\Delta G^{\text{unfolding}}$ values were obtained using alternative conformers with defined coordinates for that position. For each new reported case, SAS stability calculations were performed in different dimer chains (A, B), either in isolation or in both chains. When complex cases were studied, that is, two SASs were observed in the same patient, $\Delta G^{\text{unfolding}}$ calculations were evaluated with regards to one or both SASs per protein molecule, combined to each other in all possible associations and with dimers in active and inactive conformations. To complete the evaluation of the functional impact of each reported variant, these were also analyzed using PROVEAN (**Pro**tein **V**ariation **E**ffect **An**alyzer) and SIFT programs, and structural effects were analyzed with PolyPhen-2 (Kumar et al., 2009; Adzhubei et al., 2010; Choi et al., 2012). These methods complement FoldX as they take into account evolutionary information through sequence conservation (PROVEAN and SIFT), in addition to structural attributes in the case of PolyPhen-2. PROVEAN also allows the analysis of in-frame deletions and insertions (not possible with FoldX). Alternatively, it is possible to perform estimations for complex cases using FoldX, but this is not possible with sequence-based methods. Additionally, sequence conservation was analyzed for the reported positions. To this end sequence similarity searches were run with Blastp (Altschul et al., 1990) using the human EGFR canonical amino acid sequence as input (Universal Protein Resource, UniProtKB, P00533, isoform 1) with default settings. The searches were delimited only to EGFR molecules, duplicates and mutants were removed, and the Expected value cutoff was 10^{-4}. Corrected Shannon Entropy (CSE) was then calculated for each reported position using the server Protein Sequence Conservation Prediction (http://compbio.cs.princeton.edu/conservation/score.html) (Capra & Singh 2007).

Results

The study population which showed nonsynonymous variants in the *EGFR* gene included 135 females and 75 males, with 39% being smokers or exsmokers (7.6% did not report a smoking status). However, a reliable statistical analysis is not possible due to the low number of available cases. All novel sequence variants detected in this study are shown in Figure 1, together with their location in the human EGFR sequence/structure following EGFR sequence numbering, including the

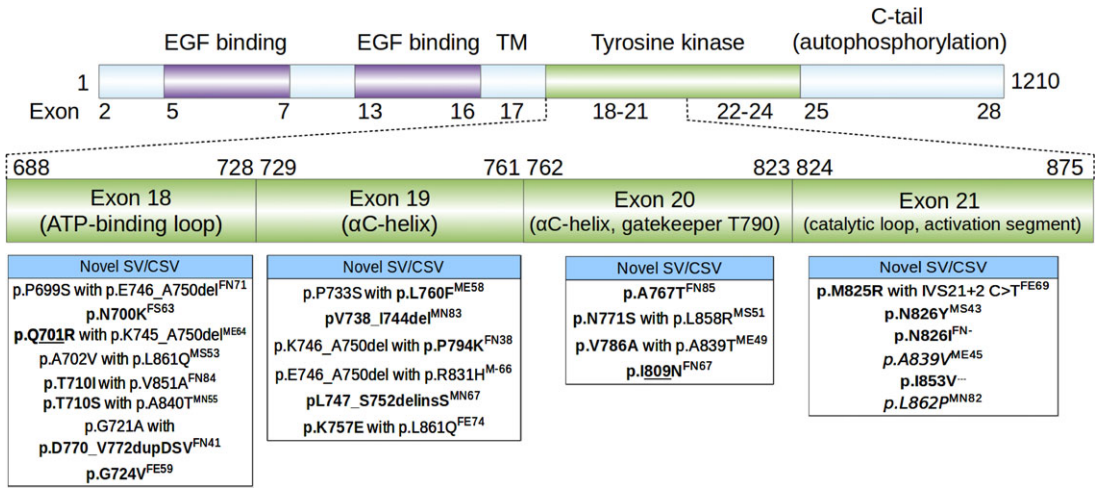

Figure 1 Distribution of novel sequence variations (SV) and combined sequence variations (CSV) in exons 18–21 of EGFR kinase domain in Argentinean NSCLC patients. Fifteen altered sites that have already been reported but changed to other different residues are bolded. Two sequence variations reported but in cancers different from NSCLC are in italic. Two altered sites never described for any type of cancer are underlined. Two combinations (CSV) not described before, but reported as individual variations, were also found. Superscripts depict information about patients as follows: F or M for gender; S, N or E for smoker, nonsmoker or exsmoker correspondingly; and a number for the age. Hyphen (-) stands for nonavailable data.

24 amino acids of the signal peptide (total length: 1210 amino acids). Figure 2 shows the reported SASs, insertion, deletion and duplication localizations in dimeric EGFR kinase domain conformations. Among the newly identified sequence variants, 18 were single amino acid substitutions (SASs), occurring alone or in combination with another sequence alteration (i.e., complex cases), one was a short deletion, one was a one short deletion-short insertion combination, and one was a duplication. Two of the SASs found in these patients were located in new positions of the EGFR kinase domain (positions 701 and 809), and had not been previously reported, even in other types of cancer. Fourteen of the new SASs were located in previously reported positions, but the change involved a different amino acid. Figure 1 also includes new and different combinations of previously reported sequence variants and those previously reported with novel cases (i.e., complex cases). The two variants in exon 19 affecting sequence length (bolded in Figure 1), a classical type of EGFR NSCLC-disease associated sequence variant, were not exactly the same as those previously reported. Also, p.A839V and p.L862P have been reported in other cancer types but, as far as we know, never in NSCLC. Each reported variant (single or complex) was found in one patient, with the exception of p.D770_P772dup, which appeared alone in one patient and together with p.Gly721Ala in another.

Structural stability calculations and functional impact results obtained with FoldX, PROVEAN, SIFT, PolyPhen-2 and CSE are included in Table 1 for monomeric and dimeric active and inactive EGFR kinase domain conformations, along with a brief description of the structural localization of the sequence variants. In the table, next to the Protein Data Bank code and pertaining to dimers, the mutated chain is indicated in each case for single variations. In complex cases, the first chain letter indicates the location of the first sequence variant and, consequently, the second. If only one chain is indicated in complex cases, both alterations belong to this chain. In the case of deletions or duplication CSE values were reported for each absent or duplicated position.

In terms of changes in structure stability, it was previously found in FoldX calculations (Juritz et al., 2012) that the best discrimination cutoff for $\Delta\Delta G$ values between polymorphic and disease-associated SASs was ± 2 kcal/mol. Taking these values into account, we classified SASs as stabilizing for those with a $\Delta\Delta G \geq 2$ kcal/mol, destabilizing for those below (-2 kcal/mol), and neutral for those with values in between. These energy estimations were performed on different forms of the EGFR protein. SASs were considered in isolation and in all possible combinations, taking into account both chains of the dimeric forms. This procedure, with 35 different combinations between structures and SASs as explained in the Materials and Methods section, allowed us to better discriminate the possible mechanistic scenarios of SASs.

Interestingly, taking into account FoldX results, differential stabilizing effects on active conformations, or a destabilizing effect on inactive conformations, were associated with a shift towards the active conformation, and, consequently, with

Table 1 *Analysis of Different Variations Found in the Reported Patients.*

		FoldX (Kcal/mol)												Disease prediction			Conservation
		Active monomer	Inactive monomer	Active dimer					Inactive dimer					PROVEAN	SIFT	PolyPhen	CSE
Region-exon. Characteristics	Variant	1M14	3W2S	4G5P_A (T790M)	4G5P_B (T790M)	4G5P_A_B (T790M)	4G5P_AB (T790M)	4G5P_BA (T790M)	2GS7_A (V948R)	2GS7_B (V958R)	2GS7_A_B (V948R)	2GS7_AB (V948R)	2GS7_BA (V948R)	Prediction	Prediction	Prediction	
JM-18, PreαC-19. New combination of two previously reported sequence variations	p.Pro699Ser	1,83	missing	1,36	0,12	1,48			missing	***-0,02***	missing			D	D	PossD	0,92
	p.Glu746_Ala750del													D			0,88 – 0,88 – 0,85 – 0,81 - 0,82
	p.Pro699Ser, p.Glu746_Ala750del																
JM-18. New SAS in a previoulsy reported altered position	p.Asn700Lys	0,73	missing	0,91	-0,23	0,42			missing	***0,09***	missing			D	D	ProbD	0,90
JM-18, PreαC-19. New SAS combined with previously reported deletion	p.Gln701Arg	0,12	-0,05	-0,37	-0,12	-0,49			***0,18***	***-0,14***	***0,07***			D	D	ProbD	0,84
	p.Lys745_Ala750del													D			0,93 - 0,88 – 0,88 – 0,85 – 0,81 - 0,82
	p.Gln701Arg, p.Lys745_Ala750del																
JM-18, Act.segm.-21. New combination of two previously reported SAS	p.Ala702Val	0,64	0,50	1,85	1,74	2,82			***0,03***	***-0,52***	***0,23***			D	T	PossD	0,78
	p.Leu861Gln	0,49	2,47	1,05	1,09	1,82			3,07	2,52	5,92			D	D	ProbD	0,96
	p.Ala702Val, p.Leu861Gln	1,20	2,91	3,01	2,43	5,51	2,92	3,18	***1,98***	***4,81***	***7,08***	***2,36***	***3,39***				
JM-18, Pre act. Segm-19. New combination of one new SAS in a previously altered position with a known SAS	p.Thr710Ile	-1,12	-0,80	-0,69	-0,71	-1,39			-1,42	-0,67	-2,24			D	D	ProbD	0,84
	p.Val851Ala	2,92	2,78	2,96	3,03	5,97			2,71	2,34	5,06			D	D	ProbD	0,87
	p.Thr710Ile, p.Val851Ala	1,83	1,95	1,94	2,25	4,52	2,12	2,12	1,29	1,67	2,82	0,86	2,11				
JM-18, Cat. Loop-21. New combination of one new SAS in a previously altered position with a known SAS	p.Thr710Ser	-0,47	-0,06	-0,23	-0,16	-0,50			-0,63	0,21	-0,56			D	T	B	0,87
	p.Ala840Thr	2,15	2,44	-0,63	1,15	0,49			2,91	2,01	4,84			D	D	ProbD	0,98
	p.Thr710Ser, p.Ala840Thr	1,68	1,52	-0,53	0,99	1,34	0,91	-0,84	2,28	2,21	4,15	1,41	2,78				
Gly-rich loop-18, PostαC-20. New combination of a previously reported SAS with a new duplication	p.Gly721Ala	4,31	3,14	2,44	9,99E-04	2,72			1,17	0,41	1,58			D	D	ProbD	0,96
	p.Asp770_Pro772dup													D			0,84 - 0,80 - 0,82
	p.Gly721Ala, p.Asp770_Pro772dup																
Gly-rich loop-18. New SAS in a previoulsy reported altered position	p.Gly724Val	17,27	13,57	24,41	8,83	32,67			9,44	13,72	23,06			D	D	ProbD	0,97
ID-19, αC-19. New combination of a new SAS in a previously reported alterated position and a previously reported SAS	p.Pro733Ser	4,00	4,35	5,21	3,96	9,21			3,62	3,90	7,58			D	D	ProbD	0,69
	p.Leu760Phe	-0,06	-0,61	-0,29	-0,33	-0,83			0,63	1,84	2,26			D	D	ProbD	0,89
	p.Pro733Ser, p.Leu760Phe	4,07	5,28	5,92	3,67	8,17	4,98	3,59	3,15	6,17	8,09	5,75	4,73				
ID-19. New deletion	p.Val738_Ile744del													D			0,89 - 0,97 - 0,99 - 0,99 - 0,99 - 0,99 - 0,99
PreαC-19, ID-20. New combination of a new SAS in a previously reported altered position and a previously reported SAS	p.Glu746_Ala750del													D			0,88 – 0,88 – 0,85 – 0,81 - 0,82
	p.Pro794Lys	1,85	0,22	1,00	3,08	4,21			1,24	1,59	3,10			D	D	ProbD	0,92
	p.Glu746_Ala750del, p.Pro794Lys																

(*Continued*)

Table 1 Continued.

PreαC-19, αE-21. New combination of two previously reported SAS	p.Glu746_Ala750del													D			0,88 – 0,88 – 0,85 – 0,81 - 0,82
	p.Arg831His	1,15	2,16	0,37	0,57	1,08			0,98	1,81	2,68			D	T	PossD	0,87
	p.Glu746_Ala750del, p.Arg831His																
PreαC-19. Previously reported combination of a deletion but with a different insertion	p.Leu747_Ser752delinsSer													D			0,88 – 0,85 – 0,81 – 0,82 – 0,85 – 0,64
αC-19 Act., Segm.-21. New combination of a new SAS in a previously reported altered position and a previously reported SAS	p.Lys757Glu	-0,15	-0,49	0,86	-0,15	0,68			-0,80	-0,49	-1,45			N	T	PossD	0,90
	p.Leu861Gln	0,49	2,47	1,05	1,09	1,82			3,07	2,52	5,92			D	D	ProbD	0,96
	p.Lys757Glu, p.Leu861Gln	0,34	1,73	2,13	0,92	2,85	1,56	1,10	2,01	2,02	3,74	1,83	2,62				
αC-19. New SAS in a previoulsy reported altered position	p.Ala767Thr	0,82	2,69	-0,72	0,63	-0,16			3,22	4,82	7,81			D	D	ProbD	0,98
PreαC-19-Act. Segm.-21. A new combination of a new SAS with a previously reported SAS	p.Asn771Ser	0,90	0,90	1,56	0,95	2,28			0,84	1,02	1,89			N	T	B	0,80
	p.Leu858Arg	-0,90	3,51	0,17	4,34	1,22			1,85	2,34	4,26			D	D	ProbD	0,95
	p.Asn771Ser, p.Leu858Arg	-0,26	4,33	1,41	3,99	5,33	5,14	1,37	2,84	3,42	6,12	3,22	2,91				
β5-20, Cat.Loop-21. New combination of a new SAS in a previously reported altered position and a previously reported SAS	p.Val786Ala	1,84	3,52	1,37	1,50	2,91			1,92	2,94	4,87			D	D	ProbD	0,81
	p.Ala839Thr	2,71	2,37	3,00	2,57	5,21			0,71	0,40	0,67			D	D	ProbD	0,99
	p.Val786Ala, p.Ala839Thr	4,81	5,67	4,09	3,07	8,36	7,38	3,28	3,01	3,34	6,02	2,39	3,89				
αDE loop-20. New altered position	p.Ile809Asn	3,82	4,22	2,59	3,01	6,18			3,51	3,26	6,52			D	D	ProbD	0,90
αE-21. New SAS in a previoulsy reported altered position	p.Met825Arg	5,43	5,04	4,72	3,75	8,51			5,01	5,77	10,20			D	D	ProbD	0,94
αE-21. New SAS in a previoulsy reported altered position	p.Asn826Tyr	-1,76	-0,08	-1,32	-1,98	-2,72			-2,07	-1,04	-3,23			D	T	PossD	0,88
αE-21. New SAS in a previoulsy reported altered position	p.Asn826Ile	-0,85	0,42	0,60	-1,17	-0,89			0,19	-0,10	0,08			D	T	PossD	0,88
Cat. Loop-21. Previously reported SAS in ovarian cancer but not in NSCLC	p.Ala839Val	3,19	3,56	0,89	2,68	4,26			2,45	2,60	4,99			D	D	ProbD	0,99
Pre act. Segm-21. New SAS in a previously reported altered position	p.Ile853Val	0,99	0,89	1,11	1,22	2,34			1,12	1,20	2,39			N	T	PossD	0,95
Act. Segm.-21. Previously reported SAS in mouth cancer but not in NSCLC	p.Leu862Pro	2,11	6,40	missing	missing	missing			5,50	4,64	10,98			D	D	ProbD	0,92

Notes: FoldX $\Delta\Delta G$ free energy calculations, PROVEAN, SIFT, and PolyPhen predictions of sequence variation impact in protein function for the Argentinean novel variants in active and inactive monomers and dimers, and amino acid conservation for each site, measured using a Corrected Shannon Entropy (CSE) score. FoldX $\Delta\Delta G$ values are expressed in Kcal/mol and rounded to two decimal places. Values higher or equal to 2 Kcal/mol and lower or equal to −2 Kcal/mol are highlighted in red and blue, respectively. Some positions have no coordinates in the PDB structures (missing). Positions P699, N700, Q701, A702 are missing in 2GS7, but we were able to calculate the corresponding $\Delta\Delta G$ using an alternative conformer (PDB 3GT8, inactive dimer), where these positions were not missing (bolded and italic $\Delta\Delta G$). PROVEAN, SIFT, and PolyPhen results are expressed according to: PROVEAN: D, deleterious; N, neutral; SIFT: D, damaging; T, Tolerated; PolyPhen: ProbD, probably damaging; PossD, possibly damaging; B, benign. Kinase regions where sequence alterations occur are abbreviated as follows: JM, juxtamembrane; PreαC, previous to αC-helix; αC, αC-helix; PostαC, after αC-helix; act. segm., activation segment; cat. loop, catalytic loop; Gly-rich, Gly-rich loop; ID, inactive dimer interface; αE, αE-helix; αED, loop between αE-helix and αD-helix; β5, β5-strand. Per-site CSE corresponds to Shannon entropy scores that have been scaled to the range [0,1] and then subtracted from 1, so that higher CSE scores indicate greater conservation.

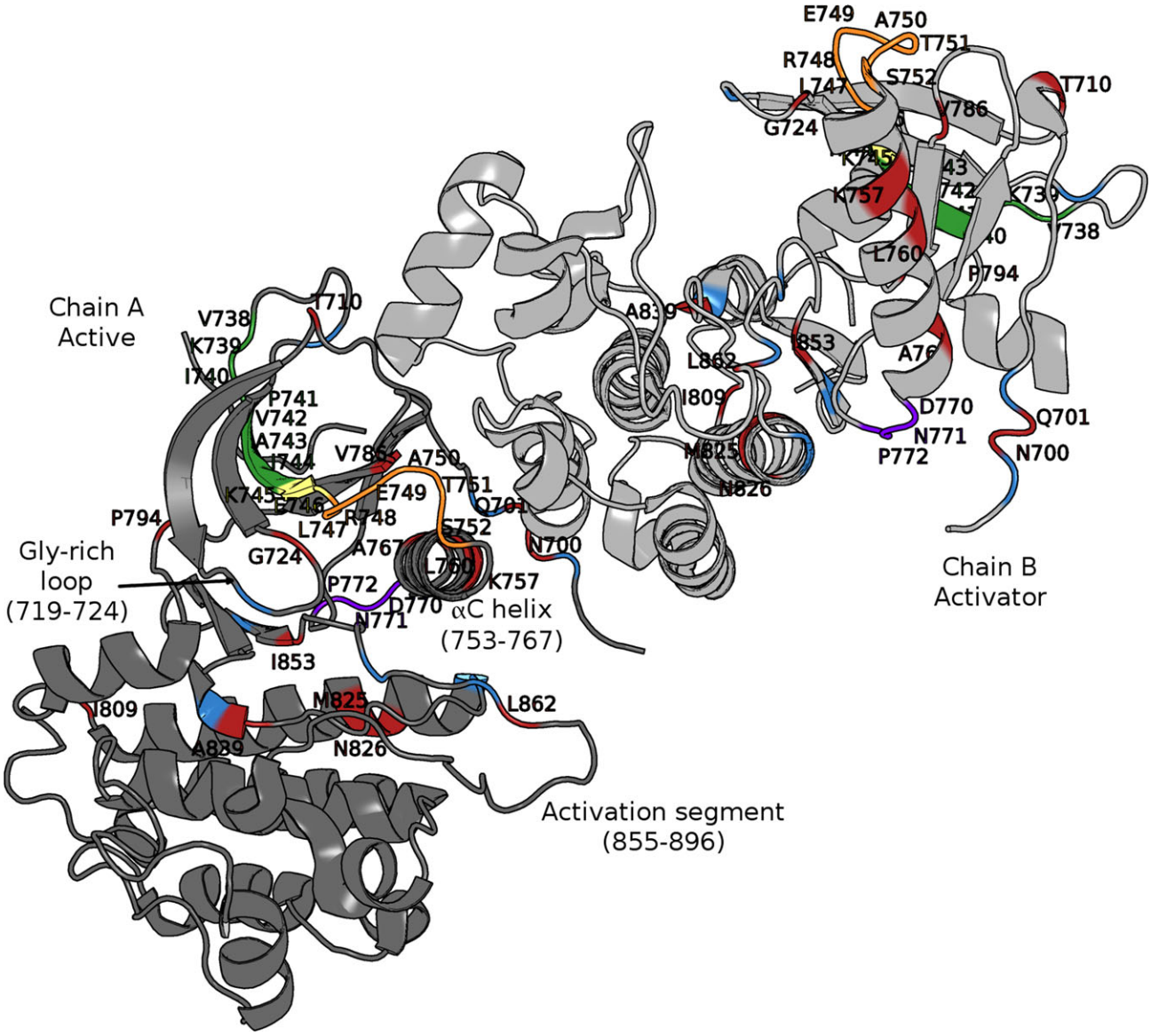

Figure 2 EGFR kinase active homodimer (PDB 4g5p) in cartoon representation. Chains A and B are in dark and light gray, respectively. Novel Argentinean SASs sites are in red and labeled using 3-letter nomenclature. V738_I744 and L747_S752 novel deletion sites are in green and orange, respectively. Between them, K745 and E746, which are part of the K745_A750 deletion previously reported, are in yellow. D770_P772 novel duplication sites are in violet. The previously reported SASs, appearing as combined sequence variants in Argentinean patients, are in light blue but not labeled.

kinase activation. In particular, 10 of the 35 combinations (~29%) differentially stabilized active conformations, while two SASs destabilized inactive conformations. As seen in Table 1, for example, the SAS p.Ala724Val shows a differential stabilizing effect on active monomeric and dimeric conformations. These effects accounted for 3.70 and 14.97 kcal/mol differences between active and inactive conformations for the monomeric and dimeric (SAS located in chain A) respectively. In the case of p.Gly721Ala, when the protein was in a monomeric conformation the SAS stabilized both conformations but with a higher value for the active form. However, when dimeric forms were analyzed, the p.Gly721Ala SAS differentially stabilized the active form while the effect was neutral for the inactive dimeric conformer. Similar explanations could be found for the rest of the activating SASs, when different structural components, such

as conformers or chains, were considered. Also, we found that p.Thr710Ile and p.Asn826Tyr differentially destabilized inactive dimeric conformations, in a way that probably favored the activation of EGFR. Consequently, the number of activating SASs reported here accounts for ~34% of cases. Alternatively, 12 of 35 combinations, p.Ala702Val+p.Leu861Gln, p.Thr710Ser+p.Ala840Thr, p.Lys757Glu+p.Leu861Gln, p.Leu760Phe, p.Ala767Thr, p.Val786Ala, p.Met825Arg, p.R831H, p.Ala840Thr, p.Leu858Arg, p.Leu858Arg+p.Asn771Ser, and p.Leu861Gln, differentially stabilized inactive conformations, representing ~34% of cases.

For the 35 combinations included in Table 1, eight (~22%) resulted in a neutral energy balance in all conformers. It is important to note that in three of these, the putative neutral effect was also supported by the results of PROVEAN, SIFT and PolyPhen-2. In two of the remaining cases the structural information was incomplete due to the presence of missing residues in all the structures evaluated for a particular conformation, although PROVEAN, SIFT and PolyPhen-2 analysis predicted the occurrence of damaging effects. Finally, in three cases FoldX neutral results were accompanied by deleterious effects according to the rest of the methods.

Activating SASs defined here using FoldX results (those stabilizing active conformations or destabilizing inactive ones) were in overall agreement with predictions by PROVEAN, SIFT, and PolyPhen-2 in ~70% of cases. However, such methods are mostly based on sequence information and do not contain conformational diversity information. So, it would be expected that some differences would be observed such as, for example, in SASs located in the 826 position: One such amino acid replacement was to Tyr (p.Asn826Tyr) and, using FoldX, we found that this had a destabilizing effect on the inactive dimeric conformation (considering the occurrence of the SAS in only one chain), which could then be classified as an activating SAS that was probably related to an increase in EGFR activity. However, the corresponding energy evaluation for another SAS in the same position, but changed to Ile (p.Asn826Ile), resulted in a neutral change. In both cases, sequence-based methods classified both SASs as neutral.

As we explained in the Materials and Methods section, stability calculations were performed by introducing one SAS per protein chain. Double SASs in the same protein molecule were also analyzed, regardless of their lower frequency compared with single sequence variations (Kobayashi et al., 2013; Tan et al., 2014). It is important to note that the analysis of complex cases showed an association with opposing stability effects, making global effect predictions almost impossible. For example, p.Val786Ala exhibited a stabilizing effect in the inactive dimeric conformer; this SAS appeared in the same tumor sample together with p.Ala839Thr that showed an active dimer stabilizing effect. Another similar case found among these patients was p.Leu760Phe combined with p.Pro733Ser.The newly reported positions, p.Gln701Arg and p.Ile809Asn, were predicted as neutral by FoldX analysis, but were predicted to be deleterious when the other analysis methods were applied. As a consequence, we can probably classify these as examples of functional SASs. Additionally, a brief comment is necessary about the well-characterized activation-segment SASs, p.Leu858Arg and p.Leu861Gln, included in the present report because these appeared in three complex cases combined with p.Asn771Ser, p.Ala702Val and p.Lys757Glu. Different research groups have established that both of these SASs stabilize the active dimeric conformation (Dixit & Verkhivker 2014); however FoldX results did not follow this conclusion, except when chain B was mutated at position 858 (the inactive conformation chain in the active dimer). However, all of these cases were predicted as deleterious using the other analysis methods. In the sequence conservation analysis, all of the reported positions show CSE values less than 1, reflecting an important degree of conservation (Table 1, CSE). Apart from the new cases, Table S1 lists all amino acid sequence variants found in the 1865 patients examined. In the same supplementary file Table 1 is included as a worksheet.

Discussion

New disease-related variant detected in different ethnic groups, along with their structural, functional and energy characterizations, could help us to understand the underlying mechanisms involved in the occurrence of both common and rare sequence variations, which would be useful in determining appropriate disease treatment strategies, as well as in the search for new drugs. The sequence variants reported here are new variants in known or novel positions, as well as in new combinations, adding new cases to this complex scenario of cancer-related EGFR variant.

The response of inhibitors in NSCLC associated with *EGFR* variant is related to the type of alteration, which, in turn, is related to changes in protein structure or function and the consequent alteration in the equilibrium between active and inactive protein conformers. The use of computational evaluations of these findings could be useful in associating new cases with old ones, as well as in finding new structure-function insights, as reported for some cases here.

It is interesting to note that several of the SASs reported in this work could differentially alter conformer stability to displace the equilibrium between altered and wild-type conformers. This finding could guide inhibitor selection in terms of the type of inhibition that would be more efficient in modulating specific conformation concentrations. However, we also found several SASs that, apparently, did not show an activating effect as predicted using FoldX estimations; instead, in most of these cases, sequence-based methods predicted damaging effects.

Overall, a combination of different methodologies enhances the predictions of the effects of SASs, as well as the corresponding mechanisms underlying their occurrence. In addition, an analysis of the resultant phenotype associated with SASs showing an opposed conformer stabilizing effect is a very interesting scenario when evaluating which SAS could dominate the progression of the illness, as well as the most appropriate TKI treatment. Unfortunately, current Argentinean routine *EGFR* allele analysis programs do not include resequencing, nor patient treatment follow-ups. Genetic variation analysis is used at the diagnostic stage to evaluate putative TKI treatment recommendations. The integration of contributions from medical, biological and protein structure/function studies could improve our current understanding of the pathological allele spectrum in this complex field.

Acknowledgements

This study was supported by the following research grants: multiyear research projects from the Consejo Nacional de Investigaciones Científicas y Técnicas (PIP CONICET), and grant 1004/11 from the Universidad Nacional de Quilmes (UNQ). The authors would like to thank the anonymous reviewers for their valuable comments and suggestions, which greatly improved the manuscript.

Conflict of Interest Statement

The authors wish to declare they have no conflicts of interest.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S. & Sunyaev, S. R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249.

Altschul, S. F., Gish, W., Miller, W., Myers E. W. & Lipman, D. J. (1990) Basic local alignment search tool. *J Mol Biol* **215**, 403–410.

Arrieta, O., Cardona, A. F., Federico, B. G., Gallo, A., Campos-Parra, A. D., Serrano, S., Castro, M., Avilés, A., Amorin, E., Kirchuk, R., Cuello, M., Borbolla, J., Riemersma, O., Becerra, H. & Rosell, R., on behalf of the CLICaP (2011) Genotyping non-small cell lung cancer (NSCLC) in Latin America. *J Thorac Oncol* **6**, 1955–1959.

Bauml, J., Mick, R., Zhang, Y., Watt, C. D., Vachani, A., Aggarwal C., Evans, T. & Langer, C. (2013) Frequency of EGFR and KRAS mutations in patients with non small cell lung cancer by racial background: do disparities exist? *Lung cancer* **81**, 347–353.

Beau-Faller, M., Prim, N., Ruppert, A. M., Nanni-Metéllus, I., Lacave, R., Lacroix, L., Escande, F., Lizard, S., Pretet, J. L., Rouquette, I., deCrémoux, P., Solassol, J., deFraipont, F., Bièche, I., Cayre, A., Favre-Guillevin, E., Tomasini, P., Wislez, M., Besse, B., Legrain, M., Voegeli, A. C., Baudrin, L., Morin, F., Zalcman, G., Quoix, E., Blons, H. & Cadranel, J. (2014) Rare EGFR exon 18 and exon 20 mutations in non-small-cell lung cancer on 10 117 patients: a multicentre observational study by the French ERMETIC-IFCT network. *Ann Oncol* **25**, 126–131.

Capra, J. A. & Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics* **23**, 1875–1882.

Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. (2012) Predicting the Functional Effect of Amino Acid Substitutions and Indels. *PLoS One* **7**, e46688.

Dixit, A. & Verkhivker, G. M. (2014) Structure-functional prediction and analysis of cancer mutation effects in protein kinases. *Comput Math Methods Med* **2014**, Article ID 653487, 24 pages.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J. W., Campbell, P. J., Stratton, M. R., & Futreal, P. A. (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic acids res* **39**(Database issue), D945–D950.

Juritz, E., Fornasari, M. S., Martelli, P. L., Fariselli, P., Casadio, R. & Parisi G. (2012) On the effect of protein conformation diversity in discriminating among neutral and disease related single amino acid substitutions. *BMC Genomics* **13**(Suppl 4), S5, 1–9.

Kobayashi, S., Canepa, H. M., Bailey, A. S., Nakayama, S., Yamaguchi, N., Goldstein, M. A., Huberman, M. S. & Costa, D. B. (2013) Compound EGFR mutations and response to EGFR tyrosine kinase inhibitors. *J Thorac Oncol* **8**, 45–51.

Kumar, P., Henikoff, S. & Ng, P. C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* **4**, 1073–1081.

Monzon, A. M., Juritz, E., Fornasari, M. S. & Parisi G. (2013) CoDNaS: a database of conformational diversity in the native state of proteins. *Bioinformatics* **29**, 2512–2514.

Pao, W. & Chmielecki, J. (2010) Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. *Nat Rev Cancer* **10**, 760–774.

Peterson, T. A., Adadey, A., Santana-Cruz, I., Sun, Y., Winder, A. & Kann, M. G. (2010) DMDM: domain mapping of disease mutations. *Bioinformatics* **26**, 2458–2459.

Roengvoraphoj, M., Tsongalis, G. J., Dragnev, K. H. & Rigas, J. R. (2013) Epidermal growth factor receptor tyrosine kinase inhibitors as initial therapy for non-small cell lung cancer: focus on epidermal growth factor receptor mutation testing and mutation-positive patients. *Cancer Treat Rev* **39**, 839–350.

Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F. & Serrano L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res* **33**(Web Server issue), W382–W388.

Sgambato, A., Casaluce, F., Maione, P., Rossi, A., Rossi, E., Napolitano, A., Palazzolo, G., Bareschino, M. A., Schettino, C., Sacco, P. C., Ciadiello, F. & Gridelli, C. (2012) The role of EGFR tyrosine kinase inhibitors in the first-line treatment of advanced non small cell lung cancer patients harboring EGFR mutation. *Curr Med Chem* **19**, 3337–3352.

Tan, D. S., Camilleri-Broët, S., Tan, E. H., Alifano, M., Lim, W. T., Bobbio, A., Zhang, S., Ng, Q. S., Ang, M. K., Iyer, N. G., Takano, A., Lim, K. H., Régnard, J. F., Tan, P. & Broët, P. (2014) Intertumor heterogeneity of non-small-cell lung carcinomas revealed by multiplexed mutation profiling and integrative genomics. *Int J Cancer* **135**, 1092–1100.

Yasuda, H., Park, E., Yun, C. H., Sng, N. J., Lucena-Araujo, A. R., Yeo, W. L., Huberman, M. S., Cohen, D. W., Nakayama, S., Ishioka, K., Yamaguchi, N., Hanna, M., Oxnard, G. R., Lathan, C. S., Moran, T., Sequist, L. V., Chaft, J. E., Riely, G. J., Arcila, M. E., Soo, R. A., Meyerson, M., Eck, M. J., Kobayashi, S. S. & Costa, D. B. (2013) Structural, biochemical, and clinical characterization of epidermal growth factor receptor (EGFR) exon 20 insertion mutations in lung cancer. *Sci Transl Med* 5, 216ra177, **1**(7), 497–514.

Zhang, Z., Stiegler, A. L., Boggon, T. J., Kobayashi, S. & Halmos, B. (2010) EGFR-mutated lung cancer: a paradigm of molecular oncology. *Oncotarget* **1**, 497–514.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Complete list of EGFR kinase domain amino acid sequence variations found in the patients with NSCLC studied in the present work.

Received: 28 December 2014
Accepted: 3 June 2015