Comparison of algorithms for the detection of cancer drivers at subgene resolution

Eduard Porta-Pardo[1,10], Atanas Kamburov[2–4], David Tamborero[5,6], Tirso Pons[7,10], Daniela Grases[1], Alfonso Valencia[8,9], Nuria Lopez-Bigas[5,6,9], Gad Getz[2–4] & Adam Godzik[1]

Understanding genetic events that lead to cancer initiation and progression remains one of the biggest challenges in cancer biology. Traditionally, most algorithms for cancer-driver identification look for genes that have more mutations than expected from the average background mutation rate. However, there is now a wide variety of methods that look for nonrandom distribution of mutations within proteins as a signal for the driving role of mutations in cancer. Here we classify and review such subgene-resolution algorithms, compare their findings on four distinct cancer data sets from The Cancer Genome Atlas and discuss how predictions from these algorithms can be interpreted in the emerging paradigms that challenge the simple dichotomy between driver and passenger genes.

Cancer is a heterogeneous disease that is driven by genomic and epigenomic abnormalities. Recent efforts in cancer genomics have provided us with a catalog detailing such abnormalities in tens of thousands of human cancers[1]. This catalog has significantly expanded our understanding of the molecular aspects of cancer. However, the mutation landscape in cancer has turned out to be extremely complex[2–4], as most tumors have hundreds or thousands of somatic mutations which are seldom found again in other tumors. This apparent heterogeneity is usually interpreted within the driver–passenger paradigm, in which the few recurrent mutations are viewed as drivers of the oncogenic process that give cancer cells a selective advantage; while most mutations, especially rare ones, are viewed as passengers that have no significant consequences for the cell[5].

There are many possible ways to identify cancer-driver events. For instance, one can look for signals of nonrandom distribution of mutations at various levels of biological resolution, spanning from individual positions in the protein[6] up to whole genes[5] or pathways[7] (**Fig. 1a**). Many of the recently developed methods aim to find driver events at the subgene level. One advantage of such higher resolution approaches is that they can identify cases when different mutations in the same gene lead to distinct phenotypes[8].

While there are several reviews of cancer-driver-detection algorithms[9,10], to the best of our knowledge none have focused on subgene-resolution algorithms, which are gaining in popularity. Here we review, classify and compare such algorithms, which we call subgene algorithms, and we discuss their strengths and weaknesses based on their results on four different cancer data sets. Note that it is not our intention to determine which methods are better, as this is something that likely depends on the type of question being asked, but rather to inform potential users about how the different assumptions and technical choices of each method influence their results. Next, we show how the results of these methods can be integrated with other biological data to gain a deeper understanding of the consequences of mutations in these driver regions. Finally, we discuss the implications that the existence of such mutation clusters might have regarding novel ideas in cancer biology, such as expanding the drivers–passenger paradigm in favor of more nuanced or even continuous models[11–13].

RESULTS

A classification of mutation-clustering algorithms

While the overall goal of all subgene driver-detection algorithms is the same—i.e., identifying nonrandom mutation clusters in cancer genomes—the details of their implementations and some of their assumptions can vary significantly. For example, some methods rely solely on protein sequences[14,15]; therefore, they can only find clusters of mutations that are linear in the primary sequence. Other methods leverage information from 3D protein structures and can identify spatial patterns that are discontinuous along the sequence[6,16]. Similarly, while some algorithms only use the position of the mutations (either in one or three dimensions) to find clusters *de novo*[16,17], others focus on externally defined protein regions (such as protein domains[18,19], phosphorylation sites[20] or interaction interfaces[21]) to identify those regions enriched in

[1]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, California, USA. [2]Department of Pathology and Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [3]Harvard Medical School, Boston, Massachusetts, USA. [4]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [5]Department of Experimental and Health Sciences, University Pompeu Fabra (UPF), Barcelona, Spain. [6]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Spain. [7]Structural Biology and BioComputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [8]Barcelona Supercomputing Centre (BSC), Barcelona, Spain. [9]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [10]Present addresses: Barcelona Supercomputing Centre (BSC), Barcelona, Spain (E.P.-P.) and Stem cells and Immunity Laboratory, National Centre for Biotechnology (CNB-CSIC), Madrid, Spain (T.P.). Correspondence should be addressed to A.G. (adam@godziklab.org).

RECEIVED 18 JANUARY; ACCEPTED 16 JUNE; PUBLISHED ONLINE 17 JULY 2017; DOI:10.1038/NMETH.4364

somatic mutations. Based on these two criteria (number of dimensions and use of externally defined regions), it is possible to classify subgene algorithms into four different groups (**Fig. 1b**,**c**). We provide an overview (detailed further in **Supplementary Table 1**) of their implementation, statistical approaches and their strengths and weaknesses.

Type I—*de novo* linear clusters

This category includes methods that look for clusters along the gene sequence. The main difference between individual methods from this group is the specific background model they use. While there are methods that rely solely on statistical models[15,22], most Type 1 methods are designed to integrate other biological signals, such as the distribution of silent mutations[14,23], the ratio between the different types of mutations occurring in a specific gene[24], the probability of each mutation given the nucleotide before and after the mutated position[25,26] or by kernel-density estimates across multiple biologically relevant scales[27].

Type II—*de novo* three-dimensional clusters

These algorithms find novel mutation clusters using information about the 3D structure of the protein encoded by a given gene. Their scope is more limited than that Type I algorithms, because Type II algorithms can only be applied to proteins whose 3D structure is either known or can be reasonably predicted. While experimentally determined structures are only available for approximately 6,100 human proteins, the structural coverage can be extended to over 13,000 proteins[28] by aligning proteins to their close homologs with experimental structures (**Supplementary Fig. 1**).

The biggest differences between Type II algorithm methods are in how such methods interpret structural data to find mutation clusters. Some tools analyze a reordered version of the protein's sequence based on the distance between residues in 3D[29] or use network algorithms on the graph derived from the structure[30]. However, most Type II algorithms are designed to identify 3D clusters using the protein structure directly and to calculate empirical *P* values by reshuffling the mutations in the structure[31]. Nevertheless, the specific details of Type II algorithms can be very different, as some use spheres of varying radii[32], while others use the closeness in the structure-derived residue network[6], the Shannon entropy of the region[17] or weighted-scoring functions[16]. Finally, while most methods can focus solely on individual proteins, others are capable of finding 3D clusters that span across protein complexes[17].

Type III—linear externally defined regions

This group contains algorithms that analyze externally defined linear protein regions to identify those that are enriched in cancer somatic mutations. Therefore, unlike Type I algorithms, these methods can only be applied to proteins where at least one functional region is known, and this currently limits their scope to approximately 65% of the human proteome (**Supplementary Fig. 1**). These regions can be protein domains[18] or post-translational modification sites[20]. Type III algorithms compare the number of mutations in the selected region with that of the rest of the protein to determine whether there is enrichment in somatic mutations in specific domains or regions. We also include in this category methods that align multiple instances of the same domain in different proteins to find commonly mutated positions[19,33].

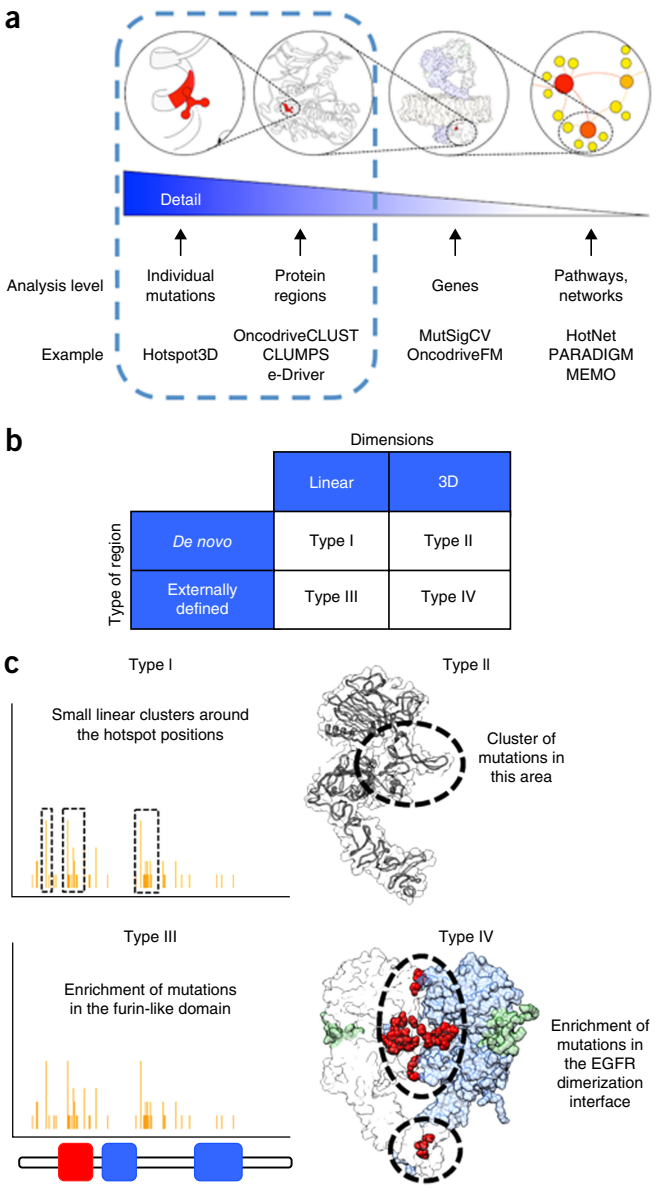

Figure 1 | Finding mutation drivers across biological scales. (**a**) Schematic of different levels for detection of cancer drivers and the tools used. (**b**) The four groups of subgene algorithms according to the type of regions they find and their number of dimensions. (**c**) Types of regions detected by each class of algorithm based on EGFR mutations in glioblastoma.

These methods are based on the rationale that mutations in equivalent positions of the same domain will have the same effect on function. These analyses have revealed strikingly similar mutation patterns across domain families such as kinases or the EGF and FGF receptor families[33].

Type IV—three-dimensional externally defined regions

Type IV algorithms find 3D externally-defined regions that are enriched in somatic mutations. To the best of our knowledge, this category currently includes only e-Driver3D[21] and a separate module of CLUMPS[16] (not used here) that uses structurally resolved interaction data. This category is most limited in scope,

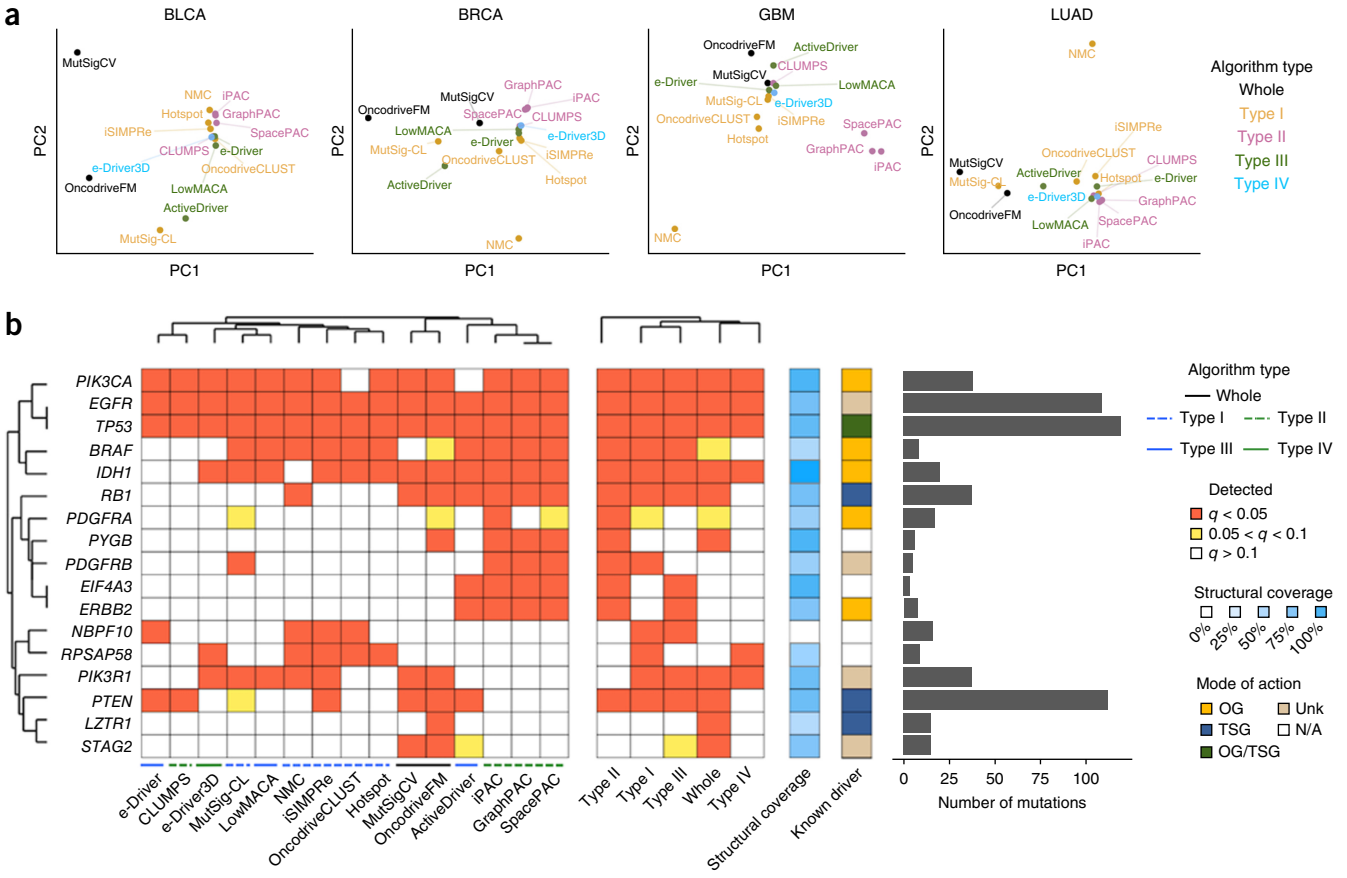

Figure 2 | Comparison of the overall predictions of each method. (**a**) Principal component analysis of the predictions by each method in the four distinct data sets using the *P* values for all the genes detected at least by one algorithm. (**b**) Predictions in the glioblastoma data set by each method (left panel) and grouped by categories (second panel). Methods are clustered according to the genes they detect. Due to space limitations, we only show genes that are either detected by at least four different algorithms or detected by a single algorithm and that are included in the Cancer Gene Census as missense drivers. We also show the structural coverage of these genes (third panel), whether they are known driver genes (fourth panel) and whether they are oncogenes (OG), tumor suppressor genes (TSG), both (OG/TSG) or known cancer genes whose mode of action still needs to be determined (Unk). Finally, we also show the number of mutations of each gene (right panel).

because both structural data and defined functional regions are required for the application of Type IV algorithms. For example, in the case of e-Driver3D, which currently analyzes protein interaction interfaces, these limitations exclude all proteins that are not involved in structurally resolved complexes. In the case of CLUMPS, the number of proteins and structures that can be analyzed is higher, as CLUMPS uses information regarding interfaces with DNA, RNA, ion ligands or small molecules in addition to protein partners. However, methods in this category exploit most biological information and, therefore, provide the highest functional information on the mutation clusters they identify.

Same category methods identify similar sets of genes

In order to explore the strengths and limitations of each of these four categories, we compared the predictions of methods covering all four categories, as well as predictions of two methods that rely on whole-gene analysis (OncodriveFM and MutSigCV)[5,34], on four different cancer genomics data sets from The Cancer Genome Atlas[1]. We aimed to assess how the specific assumptions behind each algorithm affect the number and type of drivers it identifies. In our analysis, we include five methods that belong to Type I (Hotspot[35], NMC[15], OncodriveCLUST[14], MutSig-CL[26] and iSIMPRe[22]), four from Type II (iPAC[29], GraphPAC[30], SpacePAC[32] and CLUMPS[16]), three from Type III (e-Driver[18], ActiveDriver[20] and LowMACA[19]) and one from Type IV (e-Driver3D[21]).

Our results show similarities between algorithms that belong to the same category (**Fig. 2a**, **Supplementary Figs. 2–4** and **Supplementary Tables 2–5**). For example, most Type I, Type II and Type III algorithms tend to cluster together in all data sets. Nevertheless, each group seems to have its own outlier methods. In the case of Type I algorithms, for example, NMC does not cluster with the other methods in the case of BRCA, GBM and LUAD. In the case of Type II algorithms, CLUMPS predictions are very different from those of the family of PAC algorithms in BLCA, BRCA and GBM. Finally, ActiveDriver also seems to identify different genes than the other two Type III algorithms in all data sets. The reasons why these algorithms behave differently from the rest of methods from the same category could be varied. For example, in the case of ActiveDriver, it could be because this algorithm analyzes post-translational modification sites, unlike the other two Type III algorithms, which focus on protein domains. Therefore, these tools could be finding complementary sets of genes that drive cancer through distinct mechanisms.

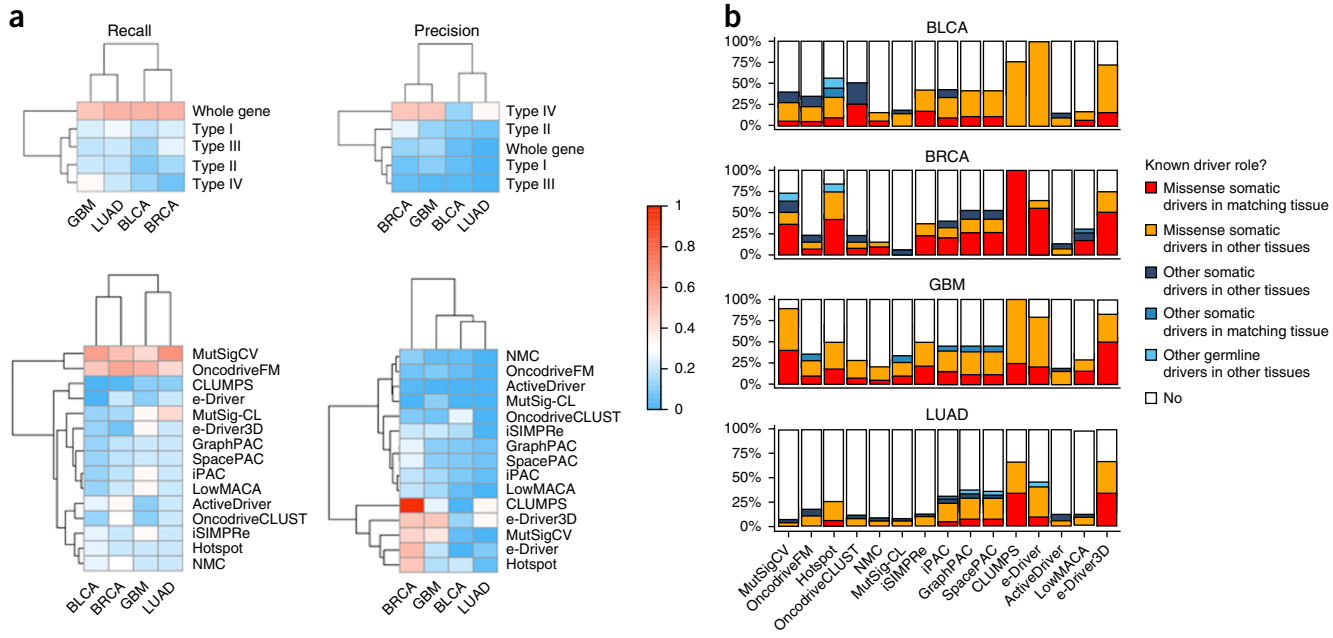

Figure 3 | Evaluating the predictions of each method and type of algorithm based on CGC data. (**a**) Recall (left) and precision (right) values for each method category in each data set (top) and each algorithm (bottom). (**b**) Known driver role of the detected genes by each method according to CGC in each data set.

In terms of specific predictions, most algorithms identify the most frequently mutated cancer driver genes in the different cancer types. For example, all methods identify *EGFR* and *TP53* as GBM driver genes, all but two find *PIK3CA*, and all but three identify *IDH1* (**Fig. 2b**). However, results for other genes exemplify the complementarity between methods from different categories. Again, in the case of GBM, Type II algorithms do not detect *PIK3R1*, because the missense mutations are spread throughout a large interface. However, Type I, Type III or Type IV algorithms detect the mutation cluster *PIK3R1*, even if the clusters predicted by each method differ slightly in its exact size and position. In other cases, certain proteins are missed by some methods simply because the methods lack statistical power at the selected significance threshold. For example, *BRAF*, a known driver gene in various cancer types, is also detected as a potential driver in glioblastoma by most subgene algorithms but, interestingly, not by the algorithms that work at the whole-gene level, OncodriveFM and MutSigCV. A possible explanation for this could be the low mutation frequency of *BRAF* in this cancer type (eight mutations in 363 samples), which makes it difficult to detect when comparing its frequency to that of other genes. Nevertheless, six of these eight mutations happen in the residue V600, making the cluster of mutations in *BRAF* amenable for detection with various subgene algorithms. In fact, many genes detected only by subgene algorithms, regardless of their category, have relatively low mutation frequencies when compared with genes identified by whole-gene algorithms (**Fig. 2b**).

Structure-based methods have high precision but low recall

We estimated the precision and recall values for each method and category in each data set using the list of genes from the Cancer Gene Census (CGC)[36] that are known to play a driver role in each cancer type. The overall results per category show that whole-gene methods have higher recall than any of the subgene categories in all four data sets we studied (**Fig. 3a**). This supports the idea that the whole-gene methods capture classical driver genes. In terms of precision, however, whole-gene methods show similar or lower values than those of the structure-based algorithms (Types II and IV).

As for the individual methods, we observe a clear split in recall values between the two whole-gene methods and any of the subgene algorithms, with the former having higher recall values than those of the latter. In our opinion, there are likely two explanations for this result. The first is that whole-gene algorithms detect both tumor suppressors and oncogenes, whereas subgene algorithms are more likely to detect oncogenes (see below). The second is that, when compared to the rest of the genome, most genes in the gold-standard list have been defined based on their mutation recurrence—the signal that whole-gene methods look for. Subgene algorithms, however, are designed to detect mutation clusters and take into account only the mutations within a specific gene. While this gives an advantage to the subgene algorithms in cases of low mutation frequencies (such as *BRAF* in glioblastoma), it is not how most cancer driver genes have been defined until now. Within subgene methods, we observe higher recall values for Type I algorithms than for the rest, probably because Type I algorithms can be applied to any gene. When analyzing the precision data, we found two groups of methods, with CLUMPS, the two e-Driver versions and Hotspot making the group with higher precision values.

Subgene methods identify new roles for known cancer genes

Intrigued by the relatively low precision values of most subgene algorithms, we classified the genes identified by each method into different categories (**Fig. 3b**) depending on whether they are known somatic drivers in that specific tissue and whether they are affected by missense mutations or through other genomic alterations (such as copy-number variations or genomic rearrangements). As expected, many of the identified genes are known to be missense somatic drivers in their corresponding tissue. However, there are also 231 genes that are predicted to be drivers by at least

one method and that, while they do not have any known driver roles in the tissue where they are detected, they are identified as drivers in other tissues. A total of 123 of these genes (53%) are missed by whole-gene methods but, nonetheless, are detected by subgene algorithms (**Supplementary Table 6**).

To our surprise, subgene algorithms also detected genes whose driver role is known but are usually affected by copy-number variations or genomic rearrangements. For example, *PDGFRB* acts as a driver in a variety of leukemias via translocations; however, iPAC, GraphPAC, SpacePAC and MutSig-CL all detected *PDGFRB* as a potential driver in GBM due to a cluster of mutations in its kinase domain. Similarly, *FGFR1* has been linked to breast cancer when amplified and to myeloproliferative syndromes when translocated. Nevertheless, both ActiveDriver and LowMACA identified a small cluster of mutations in its kinase domain. Another unexpected finding was that several genes known to cause cancer through germline (but not somatic) mutations were identified by some of the methods. The most significant example of this category is *CDK4*. Germline mutations in this gene are associated with familial melanoma, but six subgene algorithms identified *CDK4* as a likely driver in lung adenocarcinoma. Notably, some of the somatic mutations affect the same amino acids as the germline variations associated with melanoma, such as *R24L*.

Regarding the mode of action of the detected genes, it has previously been suggested that mutation clusters are more frequent in oncogenes, whereas tumor suppressor genes have more distributed mutation patterns[17] (although this notion has been questioned by recent studies[16]). Our results support the original observation, as all the subgene algorithms, regardless of the type, identify more oncogenes than tumor suppressor genes (**Supplementary Fig. 5**). In fact, when combining the predictions from all four data sets, there is a statistically significant enrichment of oncogene recognition between 1.4- and 3.7-fold in all subgene algorithms (two-sided Fisher's test; $P < 0.01$). Whole-gene methods, on the other hand, do not seem to show such bias and detect both tumor suppressor genes as well as oncogenes.

Subgene algorithms find clusters in novel driver genes

Most subgene methods identify nonrandom mutation distributions in many genes that are not part of the CGC (**Fig. 3b**). It is likely that some of these genes will be false positives, but many could be true driver genes that are missed by whole-gene methods. Just in the four cancer data sets that we studied, there are 66 genes that are not yet known to be somatic drivers and that have been detected by at least three different subgene methods but not by the methods that work at the gene resolution (**Supplementary Table 7**).

Though they are not yet part of the CGC, some of these genes have been reported to have roles in cancer or are likely to have them given their biological functions. For example, OncodriveCLUST, Hostpot and ActiveDriver all detected clusters of mutations in CSNK2A1 in lung adenocarcinoma. This protein is the catalytic subunit of the casein kinase II, a serine–threonine kinase involved in several pathways that are important in cancer, such as Wnt–CTNNB1 (ref. 37) or apoptosis[38]. Similarly, three algorithms detected a cluster of mutations in *PARP4* in the breast adenocarcinoma data set (**Supplementary Fig. 3**). Recent reports suggest that germline mutations in *PARP4* might increase the risk of thyroid cancer and breast adenocarcinoma[39], thus we believe that this gene could play an important role in these cancer types.

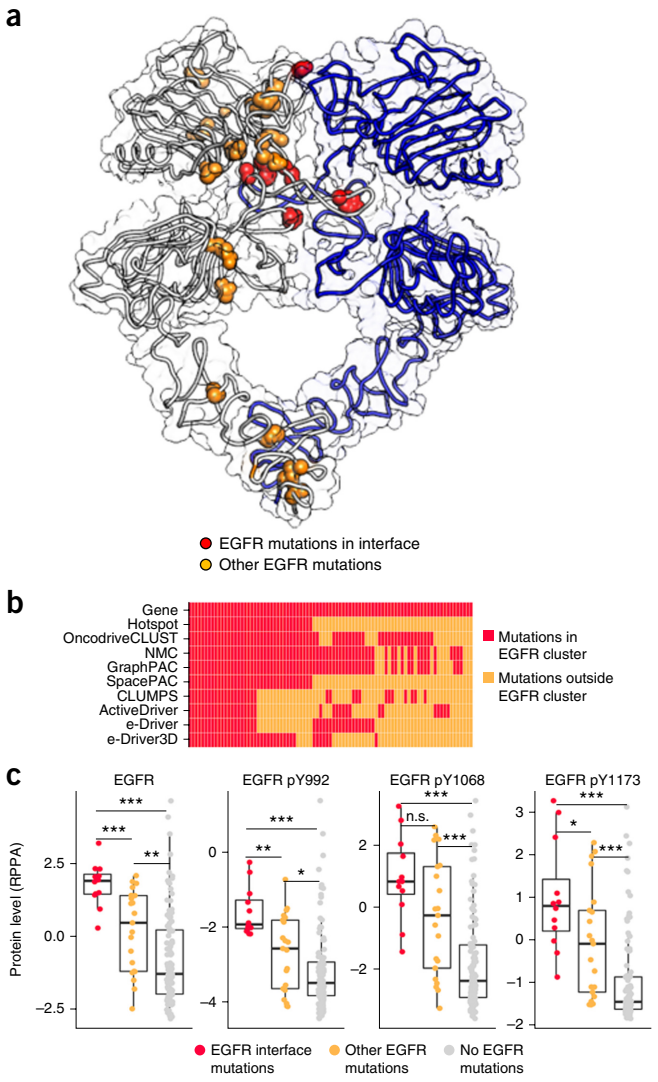

Figure 4 | Using mutation clusters to improve the definition of cancer drivers. (**a**) Glioblastoma mutations in EGFR located in the dimerization interface (in red) or in other EGFR positions (yellow). (**b**) Classification of glioblastoma samples depending on whether they have mutations in the EGFR cluster detected by each method (in red) or other EGFR mutations (in orange). Each row corresponds to a method and each column to a patient. (**c**) Comparing protein levels measured by RPPA of EGFR (left panel), EGFR pY992 (second panel), EGFR pY1068 (third panel) and EGFR pY1173 (right panel). Samples are classified according to whether they have a mutation in the EGFR–EGFR interface (in red), other EGFR mutations (in orange) or no EGFR mutations (in gray). Center, median; box limits, Q1 and Q3; high and low whiskers, highest value lower than 1.5 × Q3 and lowest value higher than 0.5 × Q1, respectively. n.s., not significant. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Overall, while one needs to be cautious when interpreting these gene lists, and further evidence is needed before the exact role of these genes and mutations in cancer is clear, we believe that subgene algorithms can identify valuable potential cancer driver genes missed by approaches that analyze mutation data at other biological scales.

Toward a continuum model of cancer mutations

Only a small number of cancer mutations have well-defined and confirmed functional consequences. Most do not, and they are usually referred to as variants of unknown significance (VUS), as

their driver effects are unknown. Many of such VUS are part of mutation clusters recognized as drivers by subgene algorithms, and this immediately raises the question whether these mutations can act as drivers in the patients that carry them. Even though it is now possible to systematically test some of these mutations experimentally[40], the most frequent approach to prioritize VUS in cancer-driver genes still involves bioinformatics tools that predict the impact of such mutations on the function of the protein[41,42] or map them into 3D structures[43,44].

Subgene algorithms also provide an obvious way for researchers to predict the impact of these variants and prioritize them. Since most of these methods identify specific clusters of positions within the protein, one can hypothesize that mutations in these positions are the most likely to be carcinogenic; whereas mutations located in other protein regions are less likely to have any significant driver effect. The power of this classification has been exemplified in the analysis of *EGFR* mutations in glioblastoma. There is a correlation between the location of *EGFR* mutations in glioblastoma and the overall level of EGFR protein as well as EGFR phosphorylation levels[21] (**Fig. 4a–c**). Samples with mutations in the dimerization interface have the highest EGFR protein and phosphorylation levels, suggesting a higher activation of the EGFR pathway, while those with mutations in other EGFR regions have an intermediate phenotype between the interface-mutated and the EGFR wild-type samples, a result that has recently been verified using cancer cell lines[6]. This is consistent with the role of interface and hotspot mutations acting as major-driver events and with other *EGFR* mutations having a different role in cancer. To the best of our knowledge, this phenomenon has not been widely studied and this is one of the few cases analyzed in more detail[16]. We believe that subgene algorithms will be key in exploring such effects.

Another important point is that the results of these algorithms can also be interpreted as an emerging challenge to the driver–passenger paradigm. Interestingly, conceptual doubts about this paradigm have been formulated for many years. For instance, it was proposed that some drivers may play a role only in specific circumstances—thus these drivers were dubbed as latent[12] or mini-drivers[13]; or they were simply considered part of a continuum of cancer-promoting mutations, each with a relatively small but additive effect[11]. Regardless of the specifics, all these models, at their core, expand on the binary driver–passenger paradigm to move toward a more nuanced classification in which mutations and the genes they affect can have different degrees of contribution to cancer growth. The results of subgene algorithms provide a natural way to classify mutations in well-established cancer-driver genes as either mutations that happen in clusters or hotspots (and more likely to be major drivers) and those that happen in other regions of the same protein and are less frequently mutated (more likely to have a lower driver effect or to even be passengers). Subgene algorithms also identify many genes that are potential low-frequency cancer-drivers; nevertheless, such genes could be important in specific cases, and the study of these genes could lead to actionable predictions as to the molecular mechanism of specific tumors they are found in.

DISCUSSION

Integrating the results of these algorithms with other omics data sets will likely have broad implications for cancer research, including, but not limited to, advancing the continuing efforts to define how mutations contribute to cancer onset and progression.

Also, while we have not explicitly explored this issue, it would likely be possible to apply the same classification (*de novo* or externally defined and linear or 3D) to algorithms that detect clusters of noncoding driver mutations. In fact, some of the algorithms discussed here have also been successfully applied to the analysis of noncoding regions[25], where they have identified several mutation clusters in promoters and 5′ UTRs, among other noncoding regions. Given the relevance of noncoding mutations in cancer[45], this will be an important issue as whole-genome sequencing becomes more widespread.

Finally, to address the issue of long-term sustainability of the benchmarking effort initiated here, we plan to incorporate the methods, input, output and gold-standard data sets into the pan-European bioinformatics infrastructure ELIXIR. ELIXIR is currently developing a data warehouse for hosting continuous automated benchmarking efforts in this and other areas of life sciences; for example, homology building, in close collaboration with different research communities. The current ELIXIR data warehouse, which includes documentation and further development plans, is accessible at http://elixir.bsc.es.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We would like to thank the people working at The Cancer Genome Atlas for their efforts and for making all the data publicly available. E.P.-P. and A.G. acknowledge the support from the Cancer Center grants P30 CA030199 (to our institute) and R35 GM118187 (A.G.). A.K. was supported by startup funds of G.G. and by a collaboration with Bayer AG. D.T. is supported by project SAF2015-74072-JIN, which is funded by the Agencia Estatal de Investigacion (AEI) and Fondo Europeo de Desarrollo Regional (FEDER). N.L.-B. acknowledges funding from the European Research Council (consolidator grant 682398). A.V. and T.P. acknowledge funding by the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 305444 (RD-Connect).

AUTHOR CONTRIBUTIONS

E.P.-P. and A.G. conceived the project. E.P.-P., D.T. and T.P. researched the data for the article. E.P.-P., A.K. and D.T. analyzed the data. All authors were involved in writing the article and reviewed and edited the manuscript before submission.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cancer Genome Atlas Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
2. Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).
3. Watson, I.R., Takahashi, K., Futreal, P.A. & Chin, L. Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.* **14**, 703–718 (2013).
4. Ortmann, C.A. *et al.* Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372**, 601–612 (2015).

5. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
6. Niu, B. *et al.* Protein-structure-guided discovery of functional mutations across 19 cancer types. *Nat. Genet.* **48**, 827–837 (2016).
7. Leiserson, M.D. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47**, 106–114 (2015).
8. Zhong, Q. *et al.* Edgetic perturbation models of human inherited disorders. *Mol. Syst. Biol.* **5**, 321 (2009).
9. Ding, L., Wendl, M.C., McMichael, J.F. & Raphael, B.J. Expanding the computational toolbox for mining cancer genomes. *Nat. Rev. Genet.* **15**, 556–570 (2014).
10. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10**, 723–729 (2013).
11. Leedham, S. & Tomlinson, I. The continuum model of selection in human tumors: general paradigm or niche product? *Cancer Res.* **72**, 3131–3134 (2012).
12. Nussinov, R. & Tsai, C.J. 'Latent drivers' expand the cancer mutational landscape. *Curr. Opin. Struct. Biol.* **32**, 25–32 (2015).
13. Castro-Giner, F., Ratcliffe, P. & Tomlinson, I. The mini-driver model of polygenic cancer evolution. *Nat. Rev. Cancer* **15**, 680–685 (2015).
14. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
15. Ye, J., Pavlicek, A., Lunney, E.A., Rejto, P.A. & Teng, C.H. Statistical method on nonrandom clustering with application to somatic mutations in cancer. *BMC Bioinformatics* **11**, 11 (2010).
16. Kamburov, A. *et al.* Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. USA* **112**, E5486–E5495 (2015).
17. Tokheim, C. *et al.* Exome-scale discovery of hotspot mutation regions in human cancer using 3D protein structure. *Cancer Res.* **76**, 3719–3731 (2016).
18. Porta-Pardo, E. & Godzik, A. e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics* **30**, 3109–3114 (2014).
19. Melloni, G.E. *et al.* LowMACA: exploiting protein family analysis for the identification of rare driver mutations in cancer. *BMC Bioinformatics* **17**, 80 (2016).
20. Reimand, J. & Bader, G.D. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.* **9**, 637 (2013).
21. Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. & Godzik, A. A pan-cancer catalogue of cancer driver protein interaction interfaces. *PLoS Comput. Biol.* **11**, e1004518 (2015).
22. Mészáros, B., Zeke, A., Reményi, A., Simon, I. & Dosztányi, Z. Systematic analysis of somatic mutations driving cancer: uncovering functional protein regions in disease development. *Biol. Direct* **11**, 23 (2016).
23. Jia, P. *et al.* MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. *Genome Biol.* **15**, 489 (2014).
24. Van den Eynden, J., Fierro, A.C., Verbeke, L.P. & Marchal, K. SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinformatics* **16**, 125 (2015).
25. Araya, C.L. *et al.* Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat. Genet.* **48**, 117–125 (2016).
26. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
27. Poole, W., Leinonen, K., Shmulevich, I., Knijnenburg, T.A. & Bernard, B. Multiscale mutation clustering algorithm identifies pan-cancer mutational clusters associated with pathway-level changes in gene expression. *PLoS Comput. Biol.* **13**, e1005347 (2017).
28. Porta-Pardo, E., Hrabe, T. & Godzik, A. Cancer3D: understanding cancer mutations through protein structures. *Nucleic Acids Res.* **43**, D968–D973 (2015).
29. Ryslik, G.A., Cheng, Y., Cheung, K.H., Modis, Y. & Zhao, H. Utilizing protein structure to identify nonrandom somatic mutations. *BMC Bioinformatics* **14**, 190 (2013).
30. Ryslik, G.A., Cheng, Y., Cheung, K.H., Modis, Y. & Zhao, H. A graph theoretic approach to utilizing protein structure to identify nonrandom somatic mutations. *BMC Bioinformatics* **15**, 86 (2014).
31. Gao, J. *et al.* 3D clusters of somatic mutations in cancer reveal numerous rare mutations as functional targets. *Genome Med.* **9**, 4 (2017).
32. Ryslik, G.A. *et al.* A spatial simulation approach to account for protein structure when identifying nonrandom somatic mutations. *BMC Bioinformatics* **15**, 231 (2014).
33. Miller, M.L. *et al.* Pan-cancer analysis of mutation hotspots in protein domains. *Cell Syst.* **1**, 197–209 (2015).
34. Gonzalez-Perez, A. & Lopez-Bigas, N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* **40**, e169 (2012).
35. Chang, M.T. *et al.* Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
36. Futreal, P.A. *et al.* A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
37. Seldin, D.C. *et al.* CK2 as a positive regulator of Wnt signalling and tumourigenesis. *Mol. Cell. Biochem.* **274**, 63–67 (2005).
38. Ahmad, K.A., Wang, G., Unger, G., Slaton, J. & Ahmed, K. Protein kinase CK2—a key suppressor of apoptosis. *Adv. Enzyme Regul.* **48**, 179–187 (2008).
39. Ikeda, Y. *et al.* Germline *PARP4* mutations in patients with primary thyroid and breast cancers. *Endocr. Relat. Cancer* **23**, 171–179 (2016).
40. Brenan, L. *et al.* Phenotypic characterization of a comprehensive set of *MAPK1*/ERK2 missense mutants. *Cell Rep.* **17**, 1171–1183 (2016).
41. Sim, N.L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.* **40**, W452–W457 (2012).
42. Creixell, P. *et al.* Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell* **163**, 202–217 (2015).
43. Mosca, R. *et al.* dSysMap: exploring the edgetic role of disease mutations. *Nat. Methods* **12**, 167–168 (2015).
44. Vázquez, M., Valencia, A. & Pons, T. Structure-PPi: a module for the annotation of cancer-related single-nucleotide variants at protein-protein interfaces. *Bioinformatics* **31**, 2397–2399 (2015).
45. Puente, X.S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).

ONLINE METHODS

Mutation data analysis and preprocessing. We compared the predictions of methods covering all four categories to explore the strengths and limitations of each of them. For Type I we used Hotspot[35], NMC[15], OncodriveCLUST[14], MutSig-CL[26] and iSIMPRe[22]. In Type II we included iPAC[29], GraphPAC[30], SpacePAC[32] and CLUMPS[16]. For Type III we included e-Driver[18,21], ActiveDriver[20] and LowMACA[19]. Finally, we used one Type IV algorithm, e-Driver3D[18,21], as well as two methods that rely on whole-gene analysis[5,12].

We analyzed four different cancer genomics data sets from The Cancer Genome Atlas: glioblastoma (GBM, n = 363)[46], breast adenocarcinoma (BRCA, n = 982)[47], bladder adenocarcinoma (BLCA, n = 137)[48] and lung adenocarcinomas (LUAD, n = 546)[49]. We used Intogen[50] to predict the location and impact of each mutation in the different protein isoforms from their genomic coordinates (**Supplementary Fig. 6**). Mutation data came from ref. 35 instead of the TCGA portal, as it had all the necessary additional information for each mutation in order to run the Hotspot algorithm.

Algorithms. We ran all algorithms using their default settings. In the case of the Hotspot algorithm, we used the genomic information of each mutation provided in the original publication. For Type II methods, when there were multiple 3D structures that could be used as templates to map the mutations, we chose the structures that had the highest structural divergence as defined by PDBFlex[51]. This limits the impact of multiple-testing issues and also ensures that we captured proteins that could be affected by protein flexibility. In the case of ActiveDriver, we used all the post-translational-modification sites provided with the algorithm: phosphorylations, acetylations and ubiquitinations. For e-Driver and e-Driver3D we used the PFAM domains, disordered regions and protein interfaces described in the original publications.

Evaluation of the results. We used the list of genes included in the Cancer Genome Census[36] (downloaded on September 12, 2016) as benchmark to compare the performance of the algorithms on known driver genes. We limited the list of genes to those that were defined as somatic and that had at least five mutations in the data set being studied. We defined a gene as predicted by an algorithm if its FDR value was below 0.05. The mode of action was also obtained from the CGC list. Note that known cancer genes that are not described as somatic (i.e., only as germline) or as drivers in other tissues in CGC are considered as not known for the purposes of the evaluation.

Regarding the PCA analysis, for each tissue we created a matrix with all the genes detected by at least two algorithms and the P values obtained by each method for each gene. For the purposes of this analysis, all the missing P values (e.g., genes with no structures have no P values for Type II or Type IV algorithms) were assumed to be 1. We calculated the PCA with the minus logarithm of the matrix. The list of candidate novel driver genes identified solely by subgene methods was limited only to those genes identified by at least three different algorithms. This threshold was defined to minimize the risk of overfitting. This approach has previously proven useful in detecting cancer driver genes[52].

EGFR RPPA analysis. We downloaded the normalized glioblastoma RPPA data from the UCSC Cancer Genome Browser[53] and compared the levels of EGFR-R-C (overall EGFR), EGFR_pY1068-R-V (EGFR phosphorylated at Y1068), EGFR_pY1173-R-C (EGFR phosphorylated at Y1173) and EGFR_pY992-R-V (EGFR phosphorylated at Y992) in three different groups of patients: those with mutations in the EGFR–EGFR interface (based on the PDB coordinates file 3NJP, chains A and B), those with other EGFR mutations and those with no mutations in EGFR. We compared protein expression levels using a two-sided Wilcoxon test.

Data availability statement. All the algorithms reviewed here can be downloaded from the sites indicated at **Supplementary Table 1**. The code and data used to compare the algorithms and generate **Figures 2–4** can be obtained at https://github.com/eduardporta/sub-gene_resolution.

46. Brennan, C.W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
47. Koboldt, D.C. *et al.* Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
48. Weinstein, J.N. *et al.* Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315–322 (2014).
49. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
50. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* **10**, 1081–1082 (2013).
51. Hrabe, T. *et al.* PDBFlex: exploring flexibility in protein structures. *Nucleic Acids Res.* **44** D1, D423–D428 (2016).
52. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
53. Goldman, M. *et al.* The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Res.* **43**, D812–D817 (2015).

doi:10.1038/nmeth.4364