cancers

MDPI

Article

Delineation of Tumor Migration Paths by Using a Bayesian Biogeographic Approach

Antonia Chroni [1,2,*], Tracy Vu [1,2], Sayaka Miura [1,2] and Sudhir Kumar [1,2,3]

[1] Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA; tug86468@temple.edu (T.V.); sayaka.miura@temple.edu (S.M.); s.kumar@temple.edu (S.K.)
[2] Department of Biology, Temple University, Philadelphia, PA 19122, USA
[3] Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[*] Correspondence: antonia.chroni@temple.edu; Tel.: +1-215-204-2938

Received: 17 October 2019; Accepted: 26 November 2019; Published: 27 November 2019

Abstract: Understanding tumor progression and metastatic potential are important in cancer biology. Metastasis is the migration and colonization of clones in secondary tissues. Here, we posit that clone migration events between tumors resemble the dispersal of individuals between distinct geographic regions. This similarity makes Bayesian biogeographic analysis suitable for inferring cancer cell migration paths. We evaluated the accuracy of a Bayesian biogeography method (BBM) in inferring metastatic patterns and compared it with the accuracy of a parsimony-based approach (metastatic and clonal history integrative analysis, MACHINA) that has been specifically developed to infer clone migration patterns among tumors. We used computer-simulated datasets in which simple to complex migration patterns were modeled. BBM and MACHINA were effective in reliably reconstructing simple migration patterns from primary tumors to metastases. However, both of them exhibited a limited ability to accurately infer complex migration paths that involve the migration of clones from one metastatic tumor to another and from metastasis to the primary tumor. Therefore, advanced computational methods are still needed for the biologically realistic tracing of migration paths and to assess the relative preponderance of different types of seeding and reseeding events during cancer progression in patients.

Keywords: biogeography; cancer; dispersal; metastasis; migration paths; tumor

1. Introduction

Cancer's uniqueness emerges from its fundamental traits (hallmarks) of tumor growth, cell expansion, and dissemination from a tumor of origin (primary) to surrounding and distant tissues (metastases) [1]. Cancer cells gain the ability to migrate, invade, and modulate tumor microenvironments. Metastasizing clones may also trigger cellular plasticity and eventually colonize secondary tissues [2]. Due to genomic instability [3–6] and the generation of intratumor genetic heterogeneity, metastatic lesions cause failures of therapeutic approaches to eradicate metastases, subsequently making metastasis the dominant cause of cancer mortality [7,8]. A comprehensive understanding of the metastatic processes is essential for cancer biology, including cancer prognosis [9] and response to treatment [10].

Metastasis involves the migration of clones (i.e., cancer cells with identical genotypes) between primary and metastatic tumor sites, both of which accumulate somatic mutations over a patient's life [11]. Thus, clones that originated from primary and metastatic tumors are evolutionarily related to each other, and their evolutionary relationship is depicted in a phylogeny. Moreover, tumor clone seeding or migration events are generally visualized in the form of migration graphs that show the relationship between the source and recipient-tumor site(s). These migration graphs can be constructed by using clone and tumor phylogenies that are frequently inferred from bulk sequencing and single-cell sequencing data [12–16].

In the past, clone seeding events were inferred by careful manual examination of cell and tumor phylogenies. However, as clone phylogenies are becoming larger due to sampling intensity (single-cell sequence data, collection of more metastases within a patient), it is expected that clone migration paths will also become more complex [17–20].

Currently, there is only one computational approach for inferring migration events between tumors (metastatic and clonal history integrative analysis, MACHINA) [21]. For a given clone phylogeny, the MACHINA method first estimates the source-tumor site as the tumor of origin for the ancestral clones at the internal nodes. A migration event is inferred whenever the location of an ancestral node in the clone tree is different from its descendant node. MACHINA uses the maximum parsimony principle in which the optimality criteria (i.e., conditions to be reached for optimum migration inferences) include: (i) the number of migration events, i.e., an event where a clone migrates from one tumor to another, (ii) the number of comigration events, i.e., an event where two separate clones from a tumor site migrate to the same tumor (polyclonal seeding event); and (iii) the number of sources of the tumor site, i.e., sources of seeding clones [21]. This approach favors solutions that minimize the number of evolutionary and migration steps involved, which may underestimate the number of sources that contribute clones to metastases. The dataset of Figure 1 is one such example where the parsimony approach taken in MACHINA failed to infer clone seeding events between metastases, as it inferred the primary tumor as the only source of tumor cells for all seeding events. The BBM method performed better than MACHINA in this example dataset, as it produced the correct migration paths.

As a result, we wondered if the limitation of parsimony criteria could be overcome by the use of Bayesian approaches that do not use such counting principles. In the field of biogeography, many methods exist for inferring the origin and movement of species/populations between (geographic) areas. Some of these methods are expected to be well suited for understanding tumor migration paths because certain biogeographic processes are analogous to tumor clone seeding and colonization events, as seen in Figure 2. More specifically, clone seeding events between distinct tumor sites could be inferred by applying biogeographic methods that model events of species/populations movement between different areas (dispersal), genetic divergence of cell lineages within an area (diversification), and the disappearance of lineage(s) from an area (extinction), as shown in Figure 2 [22,23].

In the current study, we applied a Bayesian biogeographic approach that uses an evolutionary and spatial framework for inferring species migration routes to predict migration paths between tumor sites. Specifically, we tested the Bayesian Binary MCMC (BBM) method for inferring ancestral states [24] because this method uses a full hierarchical Bayesian approach to infer dispersal, diversification, and extinction events. We also compared BBM's performance to that of MACHINA. We found that the performance of both methods is dictated by the number of tumor sites and the complexity of migration paths. Finally, we discuss the advantages of these methods and issues to be aware of when inferring tumor migration graphs.

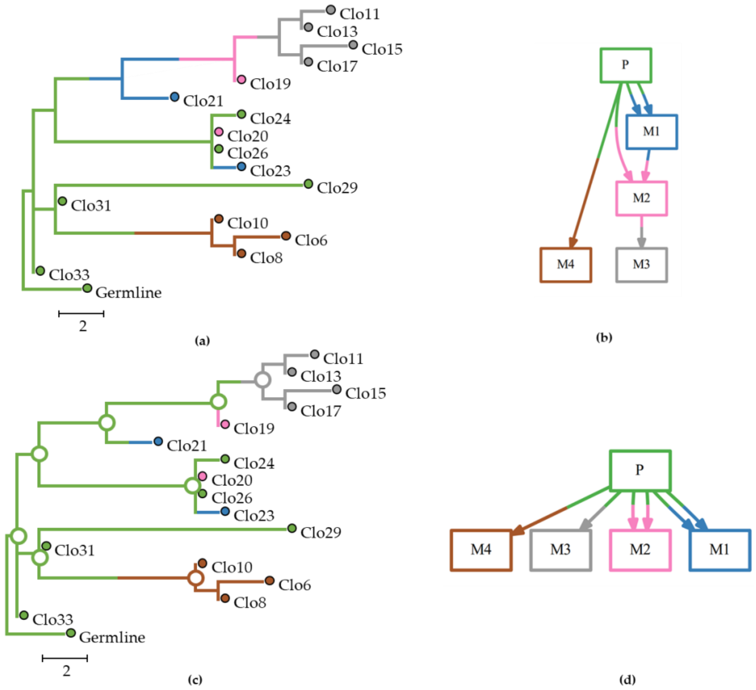

Figure 1. Analysis of an example dataset (m5Mseed76) with multisource seeding events (pM datasets) in which one tumor (M2) was seeded by clones from two other tumors (P and M1). This example dataset consisted of 16 clones and 52 characters from one primary and four metastatic tumors. Metastatic and clonal history integrative analysis (MACHINA) failed to infer the correct M1→M2 path. (**a**) The true (expected) clone phylogeny. (**b**) The true (expected) migration path. (**c**) Clone phylogeny inferred by MACHINA. (**d**) Migration graph predicted by MACHINA. Tumor clones and ancestral clones at the internal nodes are colored based on the source-tumor site: primary (green), and metastases M1 (blue), M2 (pink), M3 (gray), and M4 (brown).

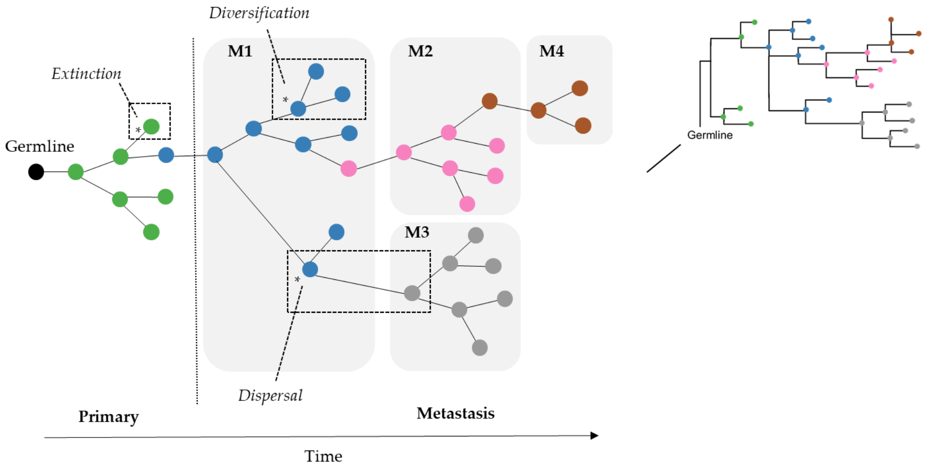

Figure 2. Cancer cells in tumors accumulate somatic mutations. Clones evolve over time within a tumor site and may go extinct or migrate to secondary tissues. In this figure, the example shows a tumor clone phylogeny with one primary and four metastatic tumor sites, each with multiple clones. Tumor migration events might be described by biogeographic processes (examples marked in dashed boxes) such as (i) dispersal: movements between areas, (ii) diversification: genetic divergence of lineages within an area, and (iii) extinction: disappearance of lineage(s) from an area. Tumor clones are colored based on the source of tumor site: primary (green), and metastases M1 (blue), M2 (pink), M3 (gray), and M4 (brown).

2. Results

We considered four different tumor clone seeding scenarios based on the numbers of seeding clones (1–3), sources of the seeding clones (primary and/or metastatic), and the presence of reseeding of the primary tumors by clone(s) from metastases. For more details see the Methods section and Figure 3. These criteria dictated the pattern and complexity of simulated migration graphs that were different for all datasets. The monoclonal Single-source (mS) seeding was the simplest scenario followed by the polyclonal Single source seeding (pS), the polyclonal Multisource seeding (pM), and the polyclonal Reseeding (pR). Our test sample contained 20 simulated datasets for each type that were further subdivided into m5 and m8 datasets based on the number of tumor sites considered (5–7 and 8–11 tumors, respectively).

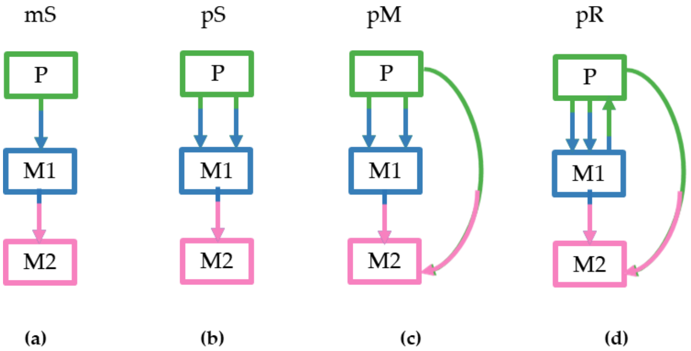

Figure 3. Examples of the four clone seeding scenarios as simulated by El-Kebir et al. [21], showing the direction of clone seeding events that were allowed to occur among tumor sites: (**a**) monoclonal single-source seeding (mS); (**b**) polyclonal single-source seeding (pS); (**c**) polyclonal multisource seeding (pM); (**d**) polyclonal reseeding (pR).

In our computational analyses, we used the Bayesian BBM method and the parsimony-based MACHINA algorithm (Parsimonious Migration History, PMH). We tested two different approaches based on the way that tree polytomies were treated in MACHINA: PMH-con, which does not attempt to resolve polytomies in the clone phylogeny, and PMH-TR, which does. More specifically, PMH-TR explores different tree topologies under which migration inferences are minimized, and jointly, refines any polytomies using the migration paths (for more details see the Methods section). For all analyzed methods, the source origin of the clone at the root of the phylogeny was assumed to be the primary tumor site. We evaluated these three approaches (BBM, PMH-con, and PMH-TR) to estimate their accuracy in inferring tumor migration paths.

2.1. *Interpretation of BBM Results in the Context of Cancer*

BBM infers ancestral distributions (locations) of species/populations at each node based on a given phylogeny and the distribution of species/populations in different locations. In cancer, clones constitute species/populations, and locations are the tumor (sampling) sites. BBM produces a source of origin of descendant clones at each ancestral node in the clone phylogeny, as well as the migration path with the highest posterior probability.

Figure 4 shows a set of multiple and reseeding clone migrations, as seen in panel a, between primary and seven metastases, and the associated clone phylogeny, as seen in panel b. In panel c, we observe the predicted tumor sites at an internal node along with their posterior probabilities produced by BBM using the clone phylogeny in panel a. Corresponding results from MACHINA are shown in panel d. Both PMH-con and PMH-TR approaches produced the same result for this dataset. The predicted ancestral tumor sites (source-tumor sites at ancestral nodes) and migration paths were congruent with the expected (true) migration paths for this example dataset and all methods, as seen in panel e.

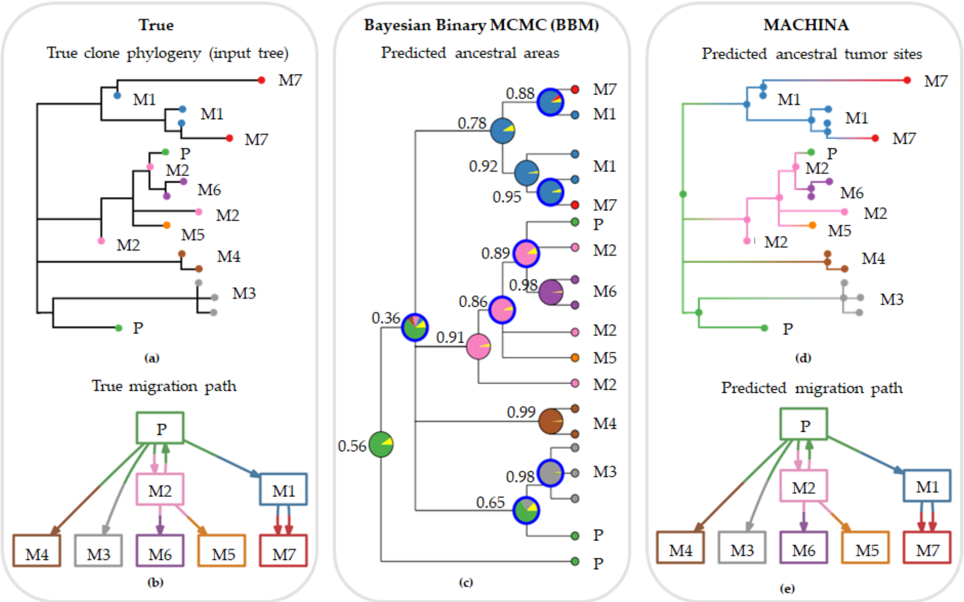

Figure 4. Inferred migration graphs for an example dataset (m8Rseed9; pR dataset). This dataset consisted of 18 clones and 66 characters from primary and seven metastatic tumors. (**a**) True (expected) clone phylogeny used as input tree for the migration inferences. (**b**) True migration path. (**c**) Ancestral areas predicted by Bayesian biogeography method (BBM) approach. Values next to the predicted ancestral ranges indicate Bayesian posterior probabilities. (**d**) Ancestral tumor sites predicted by PMH-con. PMH-TR produced the same result as PMH-con. (**e**) Predicted migration paths by MACHINA and BBM, which were consistent with the true migration graph. Tumor clones and ancestral clones at the internal nodes are colored based on the predicted source-tumor site: primary (green), and metastases M1 (blue), M2 (pink), M3 (gray), M4 (brown), M5 (orange), M6 (purple), and M7 (red). Unknown location of origin is marked as yellow in the pie charts of BBM.

2.2. *Impact of the Complexity of Migration Paths on the Inference Accuracy*

To evaluate the effect of the complexity of the dataset on the accuracy of the method, we computed F_1-scores for each dataset (see Methods). Larger values of F_1-scores indicate a more accurate inferred migration graph. We averaged the accuracy across all 20 datasets in each category and found that MACHINA and BBM to have similar overall performance, as seen in Figure 5. That is, there was not a significant accuracy difference between PMH-con and BBM (F_1 = 0.82 and 0.79, respectively), but the performance of PMH-TR was consistently worse (F_1 = 0.71), as seen in Figure 5. PMH-TR attempts to reduce the incorrect polyclonal seeding events inferred by PMH-con by jointly resolving polytomies and inferring migration paths. However, this feature introduces more errors in the migration inferences than PMH-con, most likely because PMH-TR often resolves polytomies incorrectly, as seen in Figure 5.

We next examined the impact of the increasing complexity in the history of metastatic tumor evolution, i.e., the presence of polyclonal seeding events (pS), the presence of multiple source tumor sites for seeding a metastatic tumor (pM), and the presence of reseeding event (pR). The inference of correct metastatic patterns is clearly a function of the complexity of the migration graphs, as the accuracy decreases with increasing complexity, as seen in Figure 6a. Migration paths in datasets with monoclonal seeding events (mS) were the easiest to reconstruct correctly, with overall accuracy ranging from 0.84–0.92. In contrast, the presence of polyclonal reseeding (pR) made the inference of metastatic patterns very challenging, as the accuracy declined to 0.58–0.76.

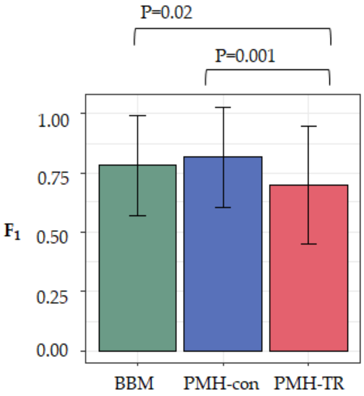

Figure 5. Overall performance based on F_1-scores of the three evaluated approaches (BBM, PMH-con, and PMH-TR) on simulated datasets.

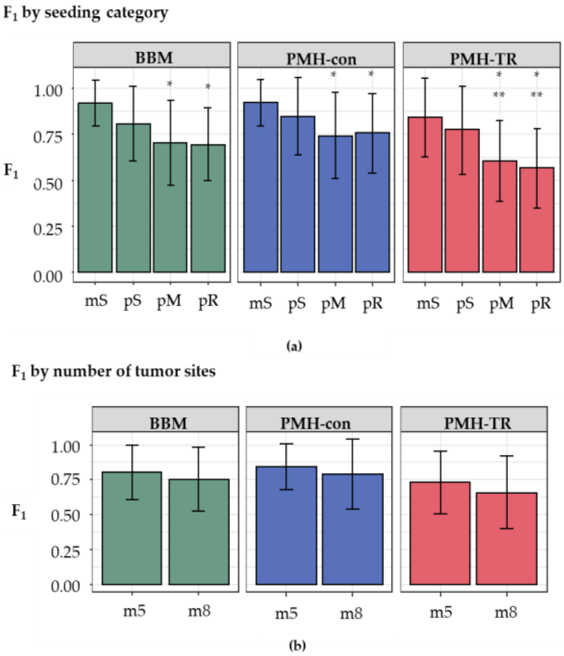

Figure 6. Overall performance (F_1-score, which is the harmonic mean of precision and recall) of different approaches for (**a**) increasing complexity of migration paths and (**b**) different number of tumor sites. Complexity of the migration graph for different seeding types was in the order: mS < pS < pM < pR. The number of tumor sites were categorized as m5 and m8 datasets (5–7, and 8–11 tumors, respectively). Differences in values of F_1-score on seeding category and tumor count were examined through *t*-test and are marked when significant (*: $p < 0.05$ with mS; **: $p < 0.05$ with pS).

Next, we conducted statistical tests to compare F_1-scores across complexity classes, as seen in Table S1. The null hypothesis of equal effect between different (simple or complex) migration schemes on the accuracy was rejected for eight out of 18 pairs of seeding scenarios at $p < 0.01$. Interestingly, a comparison of the accuracy for mS with pM (BBM and PMH-con) and pR scenarios (BBM, PMH-con, and PMH-TR) showed no significant impact of the complexity. The difference of F_1-scores between BBM and PMH-con was also not substantial in any clone seeding scenario, but PMH-TR always produced much lower F_1-score than BBM and PMH-con. PMH-TR was also not robust to these complexities, and F_1-scores decreased with increasing complexity. Thus, the inference of migration paths becomes hard for both BBM and MACHINA when the migration patterns are complex.

We also examined the impact of the number of tumor sites within a dataset. Datasets were grouped into two categories, those with a small number of tumors (5–7 tumors per dataset, m5 datasets) and those with a larger number of tumors (8–11 tumors per dataset, m8 datasets). Although the average F_1-scores of m5 datasets were slightly higher than those of the m8 for all methods (0.82 and 0.76 for BBM; 0.85 and 0.79 for PMH-con; and 0.74 and 0.67 for PMH-TR, respectively), as seen in Figure 6b, these differences were not statistically significant in the *t*-tests.

2.3. Accuracy for Different Types of Migration Paths

Migration paths can be classified into three categories based on the type of clone seeding from (i) primary to metastatic tumor site(s) (P→M), (ii) metastatic to another metastatic tumor site (M→M), and (iii) metastatic to primary tumor site (M→P). We assessed the proportion of inferred paths that were incorrect (false positives; FPs) and the proportion of correct paths that were not identified (false negatives; FNs), as seen in Figure 7.

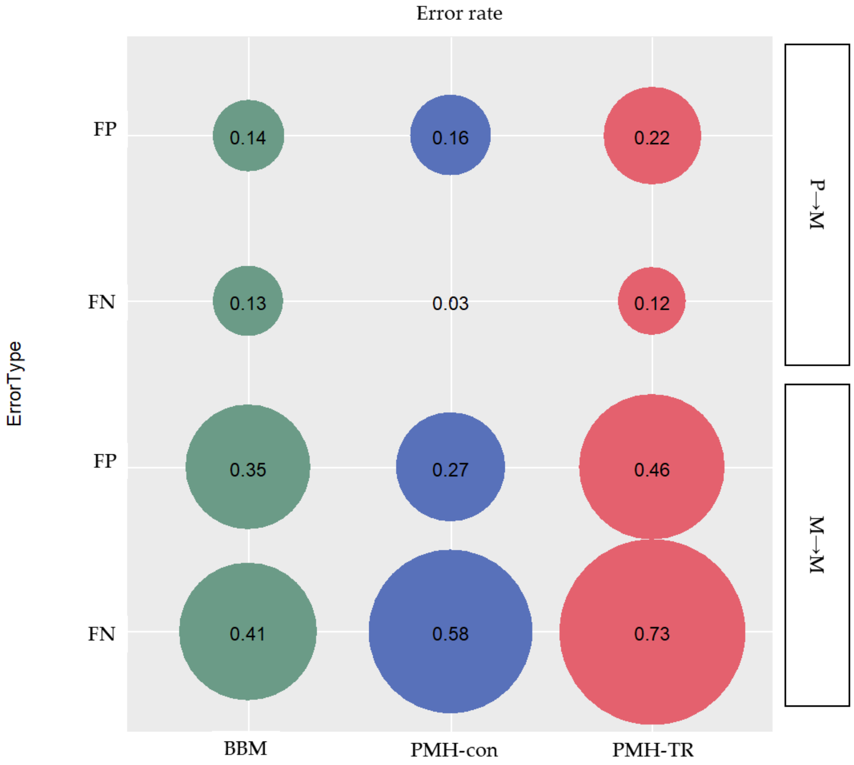

Figure 7. Error rate of incorrectly (false positives; FPs), and not identified (false negatives; FNs) inferred migration paths for the two categories of clone seeding events between different tumor sites: (i) primary and metastatic site(s) (P→M), and (ii) metastatic sites (M→M).

2.3.1. Accuracy of Migration Paths from Primary to the Metastatic Tumor Site (P→M Path)

We found that PMH-con outperformed other methods in identifying correct P→M paths. The average error rate of FNs of P→M paths was only 3%, while the errors inferred by BBM and PMH-TR were 13% and 12%, respectively, as seen in Figure 7. However, PMH-con produced more incorrect P→M paths than BBM with error rates of 16% and 14% for PMH-con and BBM, respectively. PMH-TR produced a much larger number of incorrect P→M paths than other methods, with an error rate of incorrect paths equal to 22%. That being said, the PMH-con method might have recovered more correct P→M paths than BBM, but the collection of paths inferred included many incorrect P→M paths. This pattern of performance of MACHINA may be explained by the fact that MACHINA minimizes the number of source-tumor sites in the inference of migration graphs. Consequently, MACHINA tends to generate migration schemes with a minimum number of source sites for a given dataset, e.g., one-step P→M paths will be favored over multistep P→M→M paths.

2.3.2. Accuracy of Migration Paths from Metastatic to another Metastatic Tumor Site (M→M Path)

We found that PMH-con outperformed the other methods in identifying correct P→M paths. The average error rate of FNs of P→M paths was only 3%, while the errors inferred by BBM and PMH-TR were 13% and 12%, respectively, as seen in Figure 7. However, PMH-con produced more incorrect P→M paths than BBM with error rates of 16% and 14% for PMH-con and BBM, respectively. PMH-TR produced a much larger number of incorrect P→M paths than the other methods, with an error rate of incorrect paths equal to 22%. That being said the PMH-con method might have recovered more correct P→M paths than BBM, but these include incorrect P→M paths.

2.3.3. Accuracy of Migration Paths from Metastatic to the Primary Tumor Site (M→P Path)

Similar to the inference of M→M paths, the ability to detect M→P paths was low for all of the methods (pR datasets). BBM and PMH-con produced only half of these M→P paths. Among the methods, PMH-TR performed the worst, as it rarely identified M→P paths.

Nevertheless, MACHINA did not produce any more incorrect M→P paths than BBM, most likely because MACHINA minimizes the number of source tumor sites and restricts producing migration paths that start from metastatic tumor sites. Consequently, inferences of incorrect M→P paths can be prevented more efficiently while using MACHINA than BBM. However, it is important to note that this restriction in the MACHINA method produces, more frequently, incorrect inferences of P→M paths, and less frequently, correct inferences of M→M paths, compared to BBM.

2.3.4. Overall Accuracy of Migration Paths

Overall, the BBM and MACHINA methods can detect only P→M accurately. We found that 26 and 30 inferred migration graphs among 80 datasets by BBM and PMH-con, respectively, were entirely correct, i.e., they were identical to their correct migration graphs. Around half of these datasets (11 and 15 datasets, respectively) did not contain any M→M or M→P paths. We need to be aware of this error pattern when these methods are used for empirical data analysis.

Moreover, BBM is easy to run as it is part of the RASP toolkit, but runs required ~1.5 hours, on average, and depended on the number of areas considered for the analysis (1–3.5 hours). BBM does not produce a plot of migration paths, and so, at present, it has to be drawn manually. MACHINA software is less user-friendly than BBM, as we found it to be challenging to install. It runs very fast (only a few seconds). MACHINA produces a plot of migration paths, which can be easily visualized by using software such as Graphviz (some online versions are also available, e.g., http://www.webgraphviz.com/).

3. Discussion

In this study, we evaluated a Bayesian biogeographic method as a potential alternative approach for inferring accurate cancer cell migration events between tumor sites (tumor biogeography). Although a method in biogeography has been applied to this purpose in the past [23], the accuracy of biogeographic methods is being explored for the first time in this study. Here, we applied BBM [24], because it uses a phylogenetic tree and infers dispersal patterns. We also tested the accuracy of the MACHINA method for inferring migration paths in metastasis.

Overall, we found that BBM and MACHINA produced similar results, with high accuracy in predicting a large number of migration paths correctly as long as the clone migration patterns were simple. More specifically, the complexity of datasets, i.e., presence of polyclonal seeding tumors (single source polyclonal, e.g., P→M1 and P→M1, or multiple source seeding, e.g., P→M1→M2 and P→M2) and reseeding events (M→P), impinged on the performance of the evaluated methods. On the other hand, single clone seeding events from a primary to metastasis (P→M) or between metastases (M→M) were less problematic to infer.

In more complex migration patterns between tumors, such as those with multiple source seeding in which a tumor is seeded from more than one tumor, e.g., P→M1→M2 and P→M2, the accuracy of both methods became low. The number of tumor sites and the complexity of migration graphs strongly impacted the performance of both methods. We also found that the performance of BBM excels in inferring migration paths between metastases (M→M), while MACHINA performs better for paths between primary to metastasis (P→M), and vice versa (M→P).

Each inferred migration graph generally contained one or a few incorrect migration paths. We found that most of P→M paths can be identified, but incorrect P→M paths will be additionally produced. Inferences of M→M and M→P paths are hard to identify correctly. This result could be explained by the constraint of the primary tumor as the starting tumor site for migration inferences in both MACHINA and BBM. The latter finding might indicate that the two methods are complementary to one another in validating and improving the accuracy of the metastatic inferences.

Even though BBM did not perform better than MACHINA in our direct comparisons, it is likely to be more useful because it provides more detailed information about migration paths than MACHINA. Being a Bayesian approach, BBM assigns posterior probabilities to ancestral range states for nodes in the clone phylogeny, and thus, relative probabilities of different biogeographical (dispersal, diversification, and extinction) events (if any) are produced for each ancestral node. An example is shown in Figure 4, in which the dataset included a large number of tumors, migration of multiple clones from one tumor to another, and reseeding events. In this case, the migration inferences from MACHINA and BBM were congruent with the true migration path. MACHINA produced only one possible ancestor tumor site, as seen in Figure 4d, but BBM assigns one location (or more) as a probable origin for the lineage and suggests the migration path (including probable dispersal, diversification and extinction events) for the ancestor node along with the probability for the location(s) of origin assigned at the node (for more details, see Methods section). Figure S1 shows the prediction of BBM migration paths as a result of dispersal events. For example, for the node with the assigned migration path M1→M1M7→M1 |M7 on the top of the phylogenetic tree of the Figure S1, we observe that the tumor M1 is suggested as the source of origin for the lineage. BBM further indicates that a clone from M1 migrated to tumor M7 through dispersal (event marked as blue circle around the ancestral node), where clones eventually diverged.

Examining more thoroughly the inferred migration paths in Figure S1, we also observe diversification events, denoted as, e.g., M2ˆM2. In cancer, the spread of clones is modelled through seed composition (number of clones that migrated), seed source (tumors participating in metastatic cascades), and timing (when clones diverged) [17]. BBM essentially delineates the location where genetic divergences of clones have occurred, indicating the series of tumor genetic divergence. The information on biogeographic processes inferred by BBM could be used to describe tumor clone evolutionary events in more detail. We argue that BBM is a potentially good method for modeling metastatic progression. The patterns produced by BBM may be used as a scaffold to begin to generate information on the source tumor for metastases, the number of clones involved in the initial formation of metastasis, and the route of metastatic clones among tumors in a cancer patient. This knowledge will ultimately inform the relative contribution of mutation and migration in causing intratumor heterogeneity, which is the leading cause of treatment failure and cancer mortality.

Ultimately, we must conclude that the methods available for tumor biogeography [21,23] are indeed in their infancy, as the overall accuracies are rather modest for both MACHINA and BBM. In spite of these limitations, our results clearly show promise for the application of computational methods in biogeography to infer migration graphs in cancer. By uniting the fields of population and species biogeography with tumor biogeography, we hope that it will be possible to accelerate the progress in developing more accurate computational methods for inferring migration paths between metastatic tumors. These developments will likely come from advancing the statistical framework of contemporary biogeographic methods that can integrate phylogenetic, longitudinal, and spatial signals in sequence variation. Progressively higher resolution data capturing finer details of spatiotemporal heterogeneity and evolution of tumors is likely to aid in the reliable reconstruction of migration patterns.

4. Materials and Methods

4.1. Simulated Datasets and Seeding Scenarios

To ensure a direct comparison, we used the same simulated datasets as those used to evaluate MACHINA [21]. In brief, these datasets were simulated based on the model of clone evolution and tumor growth with cell migration events modeled based on ref. [25]. Tumor migration schemes were designed and simulated based on migration (seeding) events occurring after cell cycle events (replication and death). The probabilities of clone seeding events were proportional to the number of cells in a tumor site and of the drivers considered therein. Tumor cells were allowed to migrate from an existing to another anatomical site (tumor site).

Based on the number of seeding clones (single cell, monoclonal; and group of cells, polyclonal), the number of source sites of seeding clones (single source site and multiple source site), and the presence of reseeding events (from metastatic to primary), datasets were classified into four categories. The first category of simulated tumor migration schemes included cells from one clone that could migrate from a tumor site (primary or metastatic) to another (single-source monoclonal seeding events, mS). In the second category, at least one clone per tumor site seeded another tumor site (single-source polyclonal seeding events, pS). In the third category, at least one clone per tumor site seeded two different tumor sites (multisource seeding events, pM). The fourth category contained reseeding events from metastasis to primary site, together with monoclonal seeding, single-source polyclonal seeding, and/or multisource polyclonal seeding events (reseeding, pR), as seen in Figure 3.

All the datasets were further subdivided into m5 and m8 datasets based on the number of tumor sites considered (5–7 [7] and 8–11 tumors, respectively). In total, we obtained 80 simulated datasets of clone sequences and clonal composition of primary and metastatic tumor sites, together with their histories of clone seeding events from ref. [21]. The number of primary and metastatic tumor sites ranged from 5–11. Datasets included 7–28 clones, and the number of single nucleotide variants (SNVs) was 9–99, as seen in Table S2**,** [21]. All clone phylogenies and the information of tumor sites that contain clones are available at https://github.com/raphael-group/machina/tree/master/data.

4.2. Methods for Migration Events Inference

To infer migration paths, we used the MACHINA and BBM methods. Because our interest was in testing the ability of computational methods to infer migration paths, we provided the true clone phylogenies and tumor sites. In the input tree phylogenies, branch lengths are the number of mutations, because such phylogenies are generally used on cancer research [7,26–29]. The computer-simulated data were in a binary (i.e., two-state) character system, where there were two character states for nucleotides corresponding to the normal (germline) base and to the base substitution. Mutations in the simulated sequence arose only once, meaning that the substitution process happens only towards one direction with no backward or parallel substitution changes at each nucleotide position (no homoplasy). This substitution process is in accordance with the clonal evolution model which proposes that tumors arise from a single mutated cell accumulating additional mutations [30]. When sequence data have no homoplasy, then maximum parsimony is expected to produce the correct phylogeny [31]. For that reason, we used maximum parsimony (MP) method in MEGA-CC [32] to reconstruct correct phylogenies from simulated clone sequences. All phylogenies were rooted by normal (germline) cells without any somatic mutations.

4.2.1. MACHINA Method

The Parsimonious Migration History (MACHINA) approach uses (joint) inferences of tumor clone phylogenies and/or metastatic migration histories by using DNA sequencing data [21]. We are neither examining nor discussing further the part related to clone phylogenies inferred by MACHINA, as it is beyond the scope of this study and has been discussed elsewhere [33]. We focus on the inference of migration path given the correct clone history. The MACHINA method predicts and assigns a location for each ancestral clone by using a provided list of connections between nodes and locations of tip nodes.

For MACHINA tumor migration inferences, we performed two types of analyses. In the first analysis, we constrained the ancestor at the root of the tree to be a primary tumor, and in the second, we used the Sankoff algorithm, which allowed any tumor site to be the ancestor at the root. We found that the second option produced many results, as many as 800 solutions for some datasets, as seen in Figure S2; thus, we primarily discussed findings from the constrained analysis. We also used the Parsimonious Migration History with Tree Resolution (PMH-TR) with primary tumor site constrained to examine the effect of tree polytomies in the migration inferences (analyses named as *PMH-TR*). PMH-TR resolves polytomies of a given clone phylogeny by searching for topologies that reduce the number of migration events and, subsequently, refining polytomies in a given clone tree using the migration pattern). For each approach, we allowed all possible seeding scenarios (monoclonal seeding, single/multisource polyclonal seeding, and reseeding events) to be explored. No upper bounds were placed on the number of migrations and comigrations.

4.2.2. BBM Method

To infer tumor migration patterns, we used the biogeographic Bayesian Binary MCMC (BBM) method for ancestral state [24]. BBM follows analogous transition state changes among (discrete) areas similar to nucleotide changes during DNA substitution models [34,35]. Several biogeographic methods are developed to infer migration events of species and populations with each method treating the evolutionary (cladogenesis and anagenesis events) and biogeographic (dispersal, diversification, extinction, vicariance) processes differently (for a thorough review, see ref [22]). Here, we have used a Bayesian approach, BBM, because it integrates different types of information such as phylogenetic signal in terms of posterior probabilities and/or topology (which yield Bayesian posterior probabilities for ancestor's ranges), longitudinal (divergence times), and geological/spatial (in the form of geographical coordinates or unit areas). In the current analysis, we provided the true clone phylogeny as evolutionary-type information, together with true tumor sites' locations as spatial information. This way, the same inputs were given to BBM and MACHINA.

We employed BBM analyses in RASP v4 [36]. The Markov-chain Monte Carlo chains were run three times for 5,000,000 generations to assure that MCMC chains would reach stationarity and convergence. The stationary rate frequencies and reconstructed states were sampled every 1000 generations. The first 1000 states (burn-in) were eliminated. The three independent runs were combined into a single result, which contained the inferred ancestor ranges and their probabilities. BBM ancestor ranges were estimated under the fixed Jukes–Cantor (JC69) [37] with equal dispersal rates, i.e., equal seeding rates among clones. In particular, we considered that clone seeding events between tumor sites are plausible in a migration scenario in which (a) size of the tumor site is considered not to affect dispersal patterns; (b) carrying capacities of areas and dispersal rates between tumor sites are equal; and (c) directionality of dispersal routes is allowed from any location to another [35]. The number of areas in ancestral nodes was constrained based on the number of tumor sites simulated for each dataset: (A) P, (B) M1, (C) M2, (D) M3, (E) M4, (F) M5, (G) M6, (H) M7, (I) M8, (J) M9, and (K) M10, where P, stands for primary tumor, and Mn for metastatic tumor site with n = 1–10.

The BBM method does not have the option to constrain a specific area as the ancestor at the root of the tree. To induce this constraint, we used clone phylogenies with the germline sequence as an outgroup, and we assigned the primary tumor site to be the location of the outgroup (analyses named as *BBM*).

4.3. Performance Measures

We evaluated the performances of the MACHINA and BBM methods by comparing the inferred G and simulated (expected) G* migration graphs, which were composed of migration paths, e.g., Figure 1b,d. We counted the number of correctly inferred migration paths (true positives;

TPs), those not identified (false negative; FNs), and incorrect paths (false positives; FPs). We then computed F1-score (1) for each dataset, which is the harmonic mean of precision (2) and recall (3):

$$F_1 = 2 \times \frac{\text{precisi}(1)\text{on} \times \text{recall}}{\text{precision} + \text{recall}} \tag{1}$$

where

$$\text{precision}(G, G^*) = \frac{TP}{TP + FP} \tag{2}$$

and

$$\text{recall}(G, G^*) = \frac{TP}{TP + FN} \tag{3}$$

We conducted Z-tests to assess the significance of the difference in F_1-scores between seeding scenarios and for each approach.

5. Conclusions

Currently, there is only one computational method in cancer research that infers metastatic histories. Here, we introduce a 'tumor biogeography' approach for delineating clone migration events between tumors. We have shown that overall biogeographic methods perform equally as well as MACHINA in inferring migration patterns in metastasis. Both methods produce accurate migration inferences for datasets with simple migration graphs. The performance of both methods was impacted by the complexity of datasets in terms of the number of tumor sites, presence of polyclonal seeding tumors, or multiple source seeding and reseeding events.

In conclusion, we endorse the use of biogeographic methods for inferring metastatic origin and routes as a sophisticated alternative framework. Biogeographic methods can integrate phylogenetic, longitudinal, and spatial signals of sequence data, while the current method for migration inferences between tumors cannot accommodate cancer heterogeneity in such a continuum. Progressively, more sequence data, featuring the great evolutionary and spatiotemporal heterogeneity of tumors, are becoming available. A cohort of high-accuracy that would bridge this three-dimension scale will be of high value in determining treatment strategies, appropriately designed to target the heterogeneity of metastases.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-6694/11/12/1880/s1, Figure S1: Inferred migration graph by BBM for the example dataset of Figure 4. Here the inferred migration paths due to dispersal events (blue circles around the nodes) are shown, e.g., M1→M1M7→M1 |M7. Values next to the predicted ancestral ranges indicate Bayesian posterior probabilities, Figure S2: Parsimonious Migration History (PMH) under the unconstrained mode. Inferred number of migration graphs ranged from 1–800 with the majority of the datasets to have less than 50 inferred migration graphs per input, Table S1: Comparison of the overall performances (F1-scores) between the four seeding scenarios for each analyzed approach. The z-scores are shown below the diagonal, and the corresponding P values are shown above the diagonal, Table S2: Datasets information regarding the number of anatomical (tumor) sites, clones, and SNVs included. Datasets in which a clone from a metastatic tumor site was found at the root at the rest of datasets are marked in boldface.

Author Contributions: Conceptualization, S.K. and A.C.; methodology, A.C.; formal analysis, A.C., T.V., S.M.; resources, S.K.; data curation, T.V.; writing—original draft preparation, A.C.; writing—review and editing, A.C., T.V., S.M., S.K.; supervision, S.K.; funding acquisition, S.M., S.K.

Funding: This research was funded by Temple University to S.M. (grant number: LM012758-03) and the National Institutes of Health to S.K. (grant number: LM012487-03).

Acknowledgments: We thank Jiyeong Choi, Allan George, Ryan McGregor, and Lisa Schmelkin for critical comments and technical support.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Hanahan, D.; Weinberg, R.A. Hallmarks of cancer: The next generation. *Cell* **2011**, *144*, 646–674. [CrossRef] [PubMed]
2. Welch, D.R.; Hurst, D.R. Defining the Hallmarks of Metastasis. *Cancer Res.* **2019**, *79*, 3011–3027. [CrossRef] [PubMed]
3. Bojovic, B.; Crowe, D.L. Dysfunctional telomeres promote genomic instability and metastasis in the absence of telomerase activity in oncogene induced mammary cancer. *Mol. Carcinog.* **2013**, *52*, 103–117. [CrossRef] [PubMed]
4. Vermaat, J.S.; Nijman, I.J.; Koudijs, M.J.; Gerritse, F.L.; Scherer, S.J.; Mokry, M.; Roessingh, W.M.; Lansu, N.; De Bruijn, E.; van Hillegersberg, R.; et al. Primary colorectal cancers and their subsequent hepatic metastases are genetically different: Implications for selection of patients for targeted treatment. *Clin. Cancer Res.* **2012**, *18*, 688–699. [CrossRef] [PubMed]
5. Yachida, S.; Jones, S.; Bozic, I.; Antal, T.; Leary, R.; Fu, B.; Kamiyama, M.; Hruban, R.H.; Eshleman, J.R.; Nowak, M.A.; et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **2010**, *467*, 1114–1117. [CrossRef]
6. Roschke, A.V.; Glebov, O.K.; Lababidi, S.; Gehlhaus, K.S.; Weinstein, J.N.; Kirsch, I.R. Chromosomal instability is associated with higher expression of genes implicated in epithelial-mesenchymal transition, cancer invasiveness, and metastasis and with lower expression of genes involved in cell cycle checkpoints, DNA repair, and chromatin mai. *Neoplasia* **2015**, *10*, 1222–1230. [CrossRef]
7. Hoadley, K.A.; Siegel, M.B.; Kanchi, K.L.; Miller, C.A.; Ding, L.; Zhao, W.; He, X.; Parker, J.S.; Wendl, M.C.; Fulton, R.S.; et al. Tumor evolution in two patients with basal-like breast cancer: A retrospective genomics study of multiple metastases. *PLoS Med.* **2016**, *13*, e1002174. [CrossRef]
8. Steeg, P.S. Targeting metastasis. *Nat. Rev. Cancer* **2016**, *16*, 201–218. [CrossRef]
9. Holohan, C.; van Schaeybroeck, S.; Longley, D.B.; Johnston, P.G. Cancer drug resistance: An evolving paradigm. *Nat. Rev. Cancer* **2013**, *13*, 714–726. [CrossRef]
10. Burrell, R.A.; McGranahan, N.; Bartek, J.; Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **2013**, *501*, 338–345. [CrossRef]
11. Williams, M.J.; Sottoriva, A.; Graham, T.A. Measuring clonal evolution in cancer with genomics. *Annu. Rev. Genom. Hum. Genet.* **2019**, *20*, 309–329. [CrossRef] [PubMed]
12. Deshwar, A.G.; Vembu, S.; Yung, C.K.; Jang, G.H.; Stein, L.; Morris, Q. PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **2015**, *16*, 35. [CrossRef] [PubMed]
13. Popic, V.; Salari, R.; Hajirasouliha, I.; Kashef-Haghighi, D.; West, R.B.; Batzoglou, S. Fast and scalable inference of multisample cancer lineages. *Genome Biol.* **2015**, *16*, 91. [CrossRef] [PubMed]
14. Miura, S.; Gomez, K.; Murillo, O.; Huuki, L.A.; Buturla, T.; Kumar, S. Predicting clone genotypes from tumor bulk sequencing of multiple samples. *Bioinformatics* **2018**, *34*, 4017–4026. [CrossRef]
15. Miura, S.; Huuki, L.A.; Buturla, T.; Vu, T.; Kumar, S. Computational enhancement of single-cell se- quences for inferring tumor evolution. *Bioinformatics* **2018**, *34*, i917–i926. [CrossRef]
16. Ross, E.M.; Markowetz, F. OncoNEM: Inferring tumor evolution from single-cell sequencing data. *Genome Biol.* **2016**, *17*, 69. [CrossRef]
17. Macintyre, G.; van Loo, P.; Corcoran, N.M.; Wedge, D.C.; Markowetz, F.; Hovens, C.M. How subclonal modeling is changing the metastatic paradigm. *Clin. Cancer Res.* **2017**, *23*, 630–635. [CrossRef]
18. Turajlic, S.; Swanton, C. Metastasis as an evolutionary process. *Science* **2016**, *352*, 169–175. [CrossRef]
19. Choi, Y.J.; Rhee, J.K.; Hur, S.Y.; Kim, M.S.; Lee, S.H.; Chung, Y.J.; Kim, T.M.; Lee, S.H. Intraindividual genomic heterogeneity of high-grade serous carcinoma of the ovary and clinical utility of ascitic cancer cells for mutation profiling. *J. Pathol.* **2017**, *241*, 57–66. [CrossRef]
20. Naxerova, K.; Jain, R.K. Using tumour phylogenetics to identify the roots of metastasis in humans. *Nat. Rev. Clin. Oncol.* **2015**, *12*, 258–272. [CrossRef]
21. El-Kebir, M.; Satas, G.; Raphael, B.J. Inferring parsimonious migration histories for metastatic cancers. *Nat. Genet.* **2018**, *50*, 718–726. [CrossRef] [PubMed]
22. Ronquist, F.; Sanmartín, I. Phylogenetic methods in biogeography. *Annu. Rev. Ecol. Evol. Syst.* **2011**, *42*, 441–464. [CrossRef]

23. Alves, J.M.; Prado-López, S.; Cameselle-Teijeiro, J.M.; Posada, D. Rapid evolution and biogeographic spread in a colorectal cancer. *Nat. Commun.* **2019**, *10*, 5139. [CrossRef] [PubMed]
24. Ronquist, F.; Huelsenbeck, J.P. MrBayes 3: Bayesian phylogenetic inference undermixed models. *Bioinformatics* **2003**, *19*, 1572–1574. [CrossRef] [PubMed]
25. Leung, M.L.; Davis, A.; Gao, R.; Casasent, A.; Wang, Y.; Sei, E.; Vilar, E.; Maru, D.; Kopetz, S.; Navin, N.E. Single-cell DNA sequencing reveals a latedissemination model in metastatic colorectal cancer. *Genome Res.* **2017**, *27*, 1287–1299. [CrossRef]
26. ICGC Prostate UK, Group; Gundem, G.; van Loo, P.; Kremeyer, B.; Alexandrov, L.B.; Tubio, J.M.C.; Papaemmanuil, E.; Brewer, D.S.; Kallio, H.M.L.; Högnäs, G.; et al. The evolutionary history of lethal metastatic prostate cancer. *Nature* **2015**, *520*, 353–357.
27. McPherson, A.; Roth, A.; Laks, E.; Masud, T.; Bashashati, A.; Zhang, A.W.; Ha, G.; Biele, J.; Yap, D.; Wan, A.; et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **2016**, *48*, 758–767. [CrossRef]
28. Brown, D.; Smeets, D.; Székely, B.; Larsimont, D.; Szász, A.M.; Adnet, P.-Y.; Rothé, F.; Rouas, G.; Nagy, Z.I.; Faragó, Z.; et al. Phylogenetic analysis of metastatic progression in breast cancer using somatic mutations and copy number aberrations. *Nat. Commun.* **2017**, *8*, 14944. [CrossRef]
29. Somarelli, J.A.; Ware, K.E.; Kostadinov, R.; Robinson, J.M.; Amri, H.; Abu-Asab, M.; Fourie, N.; Diogo, R.; Swofford, D.; Townsend, J.P. PhyloOncology: Understanding cancer through phylogenetic analysis. *Biochim. Biophys. Acta BBA Rev.Cancer* **2017**, *1867*, 101–108. [CrossRef]
30. Nowell, P. The clonal evolution of tumor cell populations. *Science* **1976**, *194*, 23–28. [CrossRef]
31. Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*; Oxford University Press: Oxford, NY, USA, 2000; ISBN 978-0-19-513584-8.
32. Kumar, S.; Stecher, G.; Peterson, D.; Tamura, K. Mega-cc: Computing core of molecular evolutionary genetics analysis program for automated and iterative data analysis. *Bioinformatics* **2012**, *28*, 2685–2686. [CrossRef] [PubMed]
33. Miura, S.; Vu, T.; Deng, J.; Buturla, T.; Choi, J.; Kumar, S. Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. *bioRxiv* **2019**. [CrossRef]
34. Ree, R.H.; Sanmartín, I. Prospects and challenges for parametric models in historical biogeographical inference. *J. Biogeogr.* **2009**, *36*, 1211–1220. [CrossRef]
35. Sanmartín, I.; van der Mark, P.; Ronquist, F. Inferring dispersal: A Bayesian approach to phylogeny-based island biogeography, with special reference to the Canary Islands. *J Biogeogr.* **2008**, *35*, 428–449. [CrossRef]
36. Yu, Y.; Blair, C.; He, X. RASP 4: Ancestral state reconstruction tool for multiple genes and characters. *Mol. Biol. Evol.* **2019**, msz257. [CrossRef]
37. Jukes, T.H.; Cantor, C.R. Evolution of protein molecules. *Mamm. Protein Metab.* **1969**, *3*, 132.