Recapitulating human cancer in a mouse

Lynda Chin, Frederic de Sauvage, Mikala Egeblad, Kenneth P Olive, David Tuveson & William Weiss

A panel of experts discusses the challenge of translating findings from current mouse models to the clinic.

Which mouse models are most likely to mimic the genetics and biology of a particular type and stage of human cancer is an open question. Indeed, much has been written about the failure of findings in preclinical mouse models to be translated to clinical success. At the Cold Spring Harbor Laboratory's Mechanisms and Models of Cancer meeting last year, *Nature Biotechnology* convened a panel of experts from industry and academia to debate the pros and cons of various mouse models, including GEMMs [genetically engineered mouse models], orthotopic, xenograft and allograft transplantation, with a view to improving the utility of these animals in oncology research and drug development.

What general issues should be considered when selecting a mouse model?

Kenneth P. Olive: Every type of model has its place. It depends on the question you're trying to ask. If you're trying to ask, "Does gene X function in proliferation?" then a xenograft might be a great way to look at that very quickly. If you're trying to ask, "Will drug X actually cause tumors to regress in humans?" then a different model might be more appropriate.

Lynda Chin: It also depends whether the question of interest has a relevant and appropriate model available. Defaulting to 'a model of convenience' (for example, one that is most familiar or available) will not be the most productive way to use models. For some questions, one may need to engineer new models. In fact, I think the best model will be the one that's built for purpose, designed specifically to answer a particular question, regardless of whether it uses somatic cells or is human-derived, a GEMM or a zebrafish model. The converse would mean it's unlikely for any one model to be good for everything. In the end, no single model is likely to recapitulate all aspects of the complex genetics and biology of human cancers; therefore, one has to understand the strengths and limitations of each model and leverage that.

Mikala Egeblad: Xenograft models clearly allow you to look at human-specific targets, which you cannot mimic in a GEMM. On the other hand, because the tumor microenvironment in GEMM tumor models is so much more similar to that of human tumors than are xenograft models, there may be reasons to go that route. So I agree, it really depends on what the question is.

David Tuveson: It is sobering that most of the xenograft experiments have usually ignored the tumor stroma. For example, we now realize that MET (the hepatocyte growth factor (HGF) receptor) signaling due to paracrine HGF production is a relevant resistance pathway for BRAF inhibitors in melanoma cells. We unfortunately miss this important paracrine stimulus in xenografts because mouse HGF doesn't bind well to human MET. This may be true for multiple paracrine interactions that are relevant in the tumor microenvironment, and the differential glycosylation of murine cell proteins compared to human proteins plays a role in this incompatibility. Because, for example, of the loss of stromal HGF relevance in xenografts, xenograft models may overestimate efficacy compared with GEMMs. At the same time, in GEMMs, we need to shrink the tumors, not just keep the mice alive.

How important is it to cross-validate results in different models?

K.P.O.: Proof in more models is better than proof in fewer models, but there's a flip side to that. As soon as you say we need to use lots of models to figure out what the best drug is, you've reduced your throughput and your ability to test a lot of drugs.

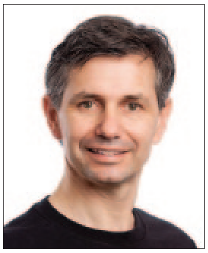
L.C.: I consider many of these preclinical studies as de-risking experiments from the perspective of clinical translation. A preclinical study

Lynda Chin is at the Department of Genomic Medicine, and the Institute for Applied Cancer Science, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; Frederic de Sauvage is at Genentech, S. San Francisco, California, USA; Mikala Egeblad is at the Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; Kenneth P. Olive is in the Departments of Medicine and Pathology and a member of the Herbert Irving Comprehensive Cancer Center at Columbia University Medical Center, New York, New York, USA; David Tuveson is at the Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA; and William Weiss is in the Departments of Neurology, Pediatrics and Neurological Surgery, University of California San Francisco, San Francisco, California, USA.
e-mail: lchin@mdanderson.org; desauvage.fred@gene.com; egeblad@cshl.edu; kenolive@columbia.edu; dtuveson@cshl.edu; waweiss@gmail.com

doesn't guarantee that the drug or mechanism would do X, Y and Z in a human. It does represent a low bar, however. You are simply saying that if the drug or mechanism doesn't pass a lower bar, the chance of it passing a higher bar is unlikely. So from that point of view, if you pass multiple challenges—to make the analogy of city champions, state champions and national champions before you go to an Olympic team—you have a better chance of making it. So I would be more excited about a drug that has cross-validation in multiple models than one that has been tested only in one.

M.E.: It is also worthwhile remembering that there may be some contexts in which a model works and some in which it doesn't. Ken and Dave, you showed in your studies, as I remember it, that subcutaneous and orthotopic xenografts, syngenic subcutaneous transplantations and GEMM models for pancreatic adenocarcinoma responded differently to the same chemotherapeutic drug [*Science* **324**, 1457–1461, 2009]. I think there is a lot of information in such experiments that compare models.

What factors are important in selecting endpoints in the models?



Frederic de Sauvage: How one measures efficacy in a particular model is very important. Of course, the claim is often made in a paper that this or that drug is great to treat cancer, but how have they measured it? Is it just delaying tumor growth? Is it stasis? Is it shrinking the tumor? That's a place where we don't really have a very good understanding of how the amplitude of the effect seen in a model—be it a xenograft, an allograft, orthotopic or a GEMM—translates into efficacy in humans.

K.P.O.: And how does the information from that mouse model relate to how the drug is used in humans? If you are looking at a primary tumor and it gets smaller, is it appropriate to initiate clinical trials in the setting of metastatic disease, where different processes may be taking place? I think in some cases, perhaps, yes, but in other cases, the mouse might be modeling slightly different processes.

F.d.S.: Yes, but a better understanding of what you can test in humans imposes stricter criteria on how you run the experiment in the preclinical model. Clearly, prevention studies in animal models are not very useful because prevention or adjuvant studies are difficult to do in humans. So people should always ensure—and there are clear published guidelines on how to run xenograft studies (for example, *Br. J. Cancer* **102**, 1555–1577, [2010])—that they consider the minimal size of the tumor when they start treatment, the minimal number of animals used, the statistics used to analyze the data and many more aspects associated with running these types of experiments.

M.E.: We need also to build models that are not just focusing on shrinkage of tumors, but on relapse and dormancy, and relapse locally and at distant sites.

What factors should be considered when using mouse data as support for clinical translation?

D.T.: With GEMMs, we should really only be looking for therapies that truly shrink established tumors. I don't think there's been enough emphasis on that point. Unfortunately, a lot of the studies that have been published with the GEMM models have been measuring the effects of drugs to prevent cancer or to shrink preneoplasms—conditions that are not pertinent for our patients who present with advanced, invasive cancers. It is frustrating when you see these papers published because now someone else has to perform an expensive and time-consuming study that the original groups should have been obligated to perform.

M.E.: Another way of thinking about this is maybe you want to use a mouse model to define what your endpoint should be in the clinical trial—turning it that way around. For example, there are cases where immune therapy doesn't immediately shrink tumors in patients; regression happens with a delay, but importantly, patients survive longer. Here, you may want to use the mouse models to try to make sure that the endpoint you use in the clinical trial represents the biology of your targeting approach and that you don't design trials that will not allow you to detect survival benefits because tumor regression doesn't occur immediately.

K.P.O.: I think there's one thing that I'd be comfortable stating unambiguously: a simple decrease in tumor growth rate in any mouse model is not what I would consider a success for the purpose of translation. Because if you put that result into a clinical context, if you take that same data and pretend it's a human being, it would be scored as 100% progressive disease and the trial would be deemed a failure. It would be a failure in the human setting, so why do we consider it a success in the mouse setting?

L.C.: I agree in general that tumor stasis in a model is not going to translate into survival benefit in patients. On the other hand, a preclinical study can be used to inform the molecular mechanism of action for a drug so that it can be rationally combined with another drug. In such cases, one can imagine that if the target of the drug is a driver of proliferation, then tumor stasis [or] stable disease can be a legitimate conclusion [or] end point of that preclinical study.

K.P.O.: Stable disease is a different thing. If you had 100% of patients in a clinical trial with stable disease in pancreatic cancer, then that would be brilliant. But I don't think you will ever be able to design a clinical trial that would be approved where the goal was to slow down the rate at which a tumor got bigger.

L.C.: I agree, but my point is that there is validity to doing a study in a model system with tumor stasis as the endpoint as long as the question being asked is something like, are gene X and drug X targeting proliferation? With such a question, seeing slowing of tumor growth in a GEMM or a xenograft model would be a legitimate conclusion; such a result can still be clinically informative (if one is considering a combination strategy). That said, personally, I don't think most mouse model studies translate immediately to human disease. Ninety-nine percent of them don't. But that doesn't mean they cannot be valuable and informative, just that our interpretations and conclusions need to be rigorous and consistent with the design of the studies. In sum, I don't think most mouse model studies should be set up solely for the purpose of immediate clinical translation.

What about selecting the timing of treatments in mouse models?

William Weiss: A lot of it has to do with where you are setting the bar, right? All things being equal, survival is a higher bar than tumor burden. Putting it another way, treating a big solid tumor represents a higher bar than injecting tumor cells with the drug and simply blocking tumor establishment.

L.C.: The reality is that with IACUC [Institutional Animal Care and Use Committee] regulation, it is difficult to measure survival because most of the animals, when their tumor grows in size or if it ulcerates, have to be euthanized. Thus, one should look at the mouse and ask, 'What advantage does it give me?' For sure, it is not the overall survival that you want for your patient. To make a mouse experiment worthwhile, it comes back to mechanism and gaining insight that you would be hard pressed to extract easily in a human experiment. In other words, you really are looking for something more when you do a mouse experiment than measurement of the tumor or overall survival.

M.E.: Before you move to a clinical trial, one of the things you can do in the mouse model is try different time points for intervention: before the tumors are detectable, when they're big and when they're metastatic disease. There was a study that Doug Hanahan and Gabriele Bergers did with MMP [matrix metalloproteinase] inhibitors and some of the angiogenesis inhibitors where they showed that different drugs had different efficacy in different phases in these mouse studies [*Nat. Cell Biol.* **2**, 737–744, 2000]. I think it is kind of sad that that wasn't translated into clinical work.

L.C.: I think it's always good to perform a long-term treatment experiment. Oftentimes a study is limited to a three-day or four-day treatment. This can be informative of mechanisms in a setting of acute treatment and response, but it is always informative if the experiment is allowed to continue for longer. Keeping the experiment going enables you to understand if you're slowing the tumor growth or if tumor growth plateaus. Or does the tumor come right back and how quickly does it come right back on the drug? It may not be relevant to a study's specific focus in terms of the mechanism, but if we increase the value of that study in terms of information that can't be a bad thing.

What issues should journals consider when reporting mouse tumor model data?

F.d.S.: It would be very useful for everyone to also report the negative data, not only the models where a drug, an siRNA [short interfering RNA] or other agent works. With the extensive amount of genomic information now available, these data would allow scientists to identify predictive markers of drug response. But to do this, we need everyone to report both the positive and the negative results, even if the experiments worked only in one out of 20 models.

L.C.: One issue that drives me nuts with many published preclinical studies is the unrealistic dosage of drugs used. Mice are given massive doses that would kill a person, or could never be achievable in a patient, or that would lose any specificity for its target.

K.P.O.: Right, because we define maximal, tolerated dose in a mouse as losing 20% of its body weight. If we look in humans, even comparatively modest side effects can serve to limit the dose.

L.C.: Even if the objective of an experiment is not direct clinical translation, I still think there should be an editorial requirement to use appropriate dosage and demonstrate that the drug hit the target in a tumor. I personally worked with [a] compound that could have potent activity in an *in vitro* kinase assay, but not in a cellular assay. In other words, it does not get into cells at all. Yet, in the literature, there are many publications reporting drug activity or tumor shrinkage by the same compound in preclinical animal models! That means that at the super high dosage given to the mouse, the compound is likely having off-target activity that is interpreted as inhibition of the intended target. Without definitively showing that the intended target is hit in a tumor *in vivo*, one cannot interpret the efficacy of a preclinical therapeutic study.

K.P.O.: Across industry, that's probably one of the biggest controls because industry is very equipped to do PK/PD [pharmacokinetic/pharmacodynamic] studies. I'd say 90% percent of the studies published in the literature in general have no PK/PD [data].

L.C.: I'm not asking for a complete PK/PD [study]. Just minimally show me that your target is in the tumor, that the drug at least got to the tumor, how about that? Right now, not only is this not cheap, it's not easy to do a complete PK/PD study in every model. So I would not want to see a publication requirement that can become inhibitory to the whole field.

F.d.S.: When thinking about what data is needed to support the claims in a paper, I would stress that not every experiment and every result needs to have immediate impact for translation. What is important is to understand that the experiment has been done correctly and what can be interpreted. If the result is that the model suggests therapeutic intervention delays growth, then that's the result, and it should be reported accurately.

K.P.O.: I would add in such a case, the last sentence of the paper should not be "this forms the basis for the translation of this drug into human patients." This happens far too often because authors are trying to hype the importance of their data.

What about heterogeneity? Do mouse models capture the heterogeneity of human tumors?

M.E.: It depends on what you mean by heterogeneity, between patients or within different regions of a single tumor?

L.C.: I think the GEMM model certainly does it better than the xenograft.

K.P.O.: But not all GEMMs are equal. I would say some that have telomerase dysfunction or that have mutant *p53* have a great deal of heterogeneity and you can see that manifest in their heterogeneity of response to therapy. There are models where if you treat them with a drug, some will have regressions, some will have stable disease, some will have tumors grow slowly and 20% of mice will progress rapidly. I think it very much depends on the model. I think you can actually use the heterogeneity of response to therapy as a tool to understand the biology. You can study what the determinants of sensitivity or resistance are within that same genetic context. For example, in a model engineered with *Kras* and *p53* mutations, you can ask, "Why did this one respond and that one not?" That is an extremely powerful approach. But I would also argue that, in general, cell line–based xenografts are going to be far less heterogeneous.

F.d.S.: Why is a *Kras* or *p53* deleted GEMM more heterogeneous than a broad panel of cell lines?

K.P.O.: Because we see it manifest in the response to therapy. From the data, you can see that there is more heterogeneity in the GEMMs. You're saying you see heterogeneity in a panel of cell lines, which is different because that's a panel of [*in vitro* cell] models. I could say I could have a panel of GEMMs. That's not a fair comparison. When compared to an individual cell line–based xenograft, the heterogeneity in a GEMM is greater because [the tumors] are undergoing an evolutionary process developing from precursor lesions, which will produce multiple lineage branches with different mutational complements. Cell lines already went through that process in their original host, but then are homogenized *in vitro*. Once they are implanted into a recipient mouse, there is little need for additional evolution and little heterogeneity available to serve as a basis for selection.

F.d.S.: You know mouse models of lung tumors are limited pretty much to a few genes such as *Kras*, *p53* or *Lkb1*. There is built-in heterogeneity. It accumulates in these models, but you can get as much heterogeneity with the hundred or so lung cell lines that are available. I wouldn't say one is necessarily better than the other at this time.

Which models are most effective at mimicking metastasis?

K.P.O.: We have mouse models that develop metastases routinely and are taken to represent metastatic disease. Some of them are very high-fidelity models, by many criteria. However, dying with metastasis and dying of metastasis are two profoundly different things. That's something I learned, I think, in my own research very recently. If you want to model the effects of a drug in the setting of metastatic cancer, you have to have a model that dies as a result of its metastatic burden. Just because the mice have some micrometastases in their liver doesn't mean that they're dying of metastatic disease. They're dying with metastatic disease.

M.E.: I think very few of our models really are modeling the clinical reality where you do whatever you can to remove the primary tumor. Most of our models just don't do that.

K.P.O.: I would advocate an investment in mouse surgical skills to remove primary tumors and then see what happens in the metastases there. It's doable. It's not something that's typically done. A similar issue is arising in the prostate cancer field; nearly all nonresectable prostate cancer patients undergo hormone ablation as a first step. Therefore, it would make sense to set up mouse treatment studies in a similar manner, waiting for the development of hormone ablation–resistant tumors before beginning treatment with a novel therapy.

F.d.S.: There would definitely be a benefit from a more organized definition of metastasis. At the moment, just showing PCR-positive cells in a tissue qualifies as metastasis. But that has little value in terms of relevance to the clinical course of the patient.

L.C.: I think more efforts should be invested in making models for metastatic progression. That hasn't been a major focus in my own lab or broadly in the community. And sometimes we all suffer from being comfortable using the model we have and trying to use it for all purposes. I keep trying to remind myself, and those in my lab, that just because you have a great model for X doesn't mean that it's any good for Y. If we really want to study metastasis in GEMM models, we need to invest in innovating and engineering new models that have the metastatic progression phenotype.

W.W.: Another aspect of metastasis associated with some cancers is that it happens early in disease. That is a process that isn't currently modeled that well in GEMMs.

How can we improve the reproducibility of published results from mouse models?

L.C.: I think one aspect that has to do with reproducibility is the robustness of the phenotype being described. Oftentimes, the observational conclusions are true in the limited context described in a publication, for example, a xenograft study using four cell lines, two *KRAS*-mutant and two *KRAS*-wild type. But the question about the robustness of that phenotype is how generalizable is that finding when you repeat the study in 20 *KRAS*-mutant or 200 *KRAS* wild-type cell lines.

F.d.S.: It also depends on how confident you are about the result. Say you discover an *ALK* translocation and show that in one or two *ALK* mutant cell lines you've got efficacy of an inhibitor against the target. It smells right. If the efficacy is really convincing and if the experiment appears to be well done, then I would say you can have great confidence about reproducibility, the correlation with the genotype and what patients to treat. In many indications, if you were going to treat a small subset of patients and you know how to select them, it's a win.

D.T.: There's been some talk about creating journals that publish papers from people that simply reproduce a previous paper and from people that can't reproduce that result. If a group does reproduce a result, probably you need only one group to show that; but if someone can't reproduce a finding, probably you'd need [at least] two groups to show that. But the reproducibility issue is just one side. I think as long as you publish high-quality, mechanistic-based findings, data from mouse models are going to be extremely, profoundly informative. They always have been. Mice are one type of model. But the best models will always be patients.

L.C.: I think there are certain standards when working with mouse models that most people would agree to. Implementing this more widely across the board would help reproducibility because these minimum standards are not enforced in many of the published studies.

K.P.O.: I would advocate that any paper that claims to be performing a true preclinical trial as a basis for bringing a drug into human beings should be put into a different category and should be held to different criteria and different standards. The reward would be a higher impact for passing those standards. That doesn't mean that a fundamental research project isn't going to get published. I'm just saying, if you're going to make a claim that you have established the rationale for a clinical trial then you should be held to a higher standard.