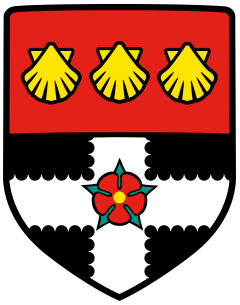
A quantitative approach to signal processing in cancer cell dispersal



George Butler

School of Biological Sciences

University of Reading

This thesis is submitted for the degree of

Doctor of Philosophy

July 2021

This thesis is dedicated to my loving parents

Martin and Lorraine Butler

who have provided unwavering support at even the most testing of times.

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

George Butler

July 2021

Publications

Chapter 3 is published as:

Butler, G., Keeton, S.J., Johnson, L.J. and Dash, P.R. (2020) *A phenotypic switch in the dispersal strategy of breast cancer cells selected for metastatic colonization*, Proceedings of the Royal Society B:Biological Sciences. doi: 10.1098/rspb.2020.2523.

Author contributions are as followed:

- G.B, S.K, L.J and P.D conceived and designed the study.
- S.K evolved the populations and collected the time-lapse data.
- G.B developed the methodology and performed the formal analysis.
- L.J and P.D supervised the work.
- G.B wrote the manuscript.

Additional publications:

Wass, A.V., Butler, G., Taylor, T.B., Dash, P.R. and Johnson, L.J. (2020) *Cancer cell lines show high heritability for motility but not generation time*, Royal Society Open Science. doi: 10.1098/rsos.191645.

Author contributions are as followed:

- A.W, P.D, T.T and L.J conceived the project.
- A.W conducted laboratory work and collected data.
- G.B and A.W conducted statistical analyses.

- P.D and L.J co-supervised the project.
- All authors contributed to writing the manuscript.

Butler, G., Rudge, J. and Dash, P. R. (2019) *Mathematical modelling of cell migration*, Essays in Biochemistry. doi: 10.1042/EBC20190020.

Author contributions are as followed:

- G.B and P.D conceived and designed the review.
- G.B wrote the initial manuscript.
- All authors contributed to subsequent drafts.

Acknowledgements

First and foremost, I would like to thank my supervisors Phil Dash and Louise Johnson for giving my the opportunity to attempt this PhD as well as the academic freedom and trust to pursue an eclectic mix of research questions and ideas. Likewise, I would also like to thank Shirley Keeton who evolved all of the cancer cell populations within this thesis and collected all of the corresponding time-lapse data. The results from any model only as good as the data upon which the model is built and thus Shirley has been instrumental in the insight gained from this thesis.

Next, I would like to thank the entire Evolution Group at the University of Reading for which I was fortunate enough to based throughout my PhD. In particular, I would like to thank Chris Venditti and Mark Pagel who have both more hugely inspiring throughout my PhD and have actively encouraged me to pursue a highly quantitative, but biological relevant, approach in all aspects of my work. Similarly, Andrew Meade has provided me with invaluable computational guidance and support especially with regards to the particularly tricky computational calculation of Zernike moments. In addition, I would also like to thank Jo Baker and Ciara O'Donovan for their comical, albeit unintentional, antics throughout my PhD that have provided continuous amusement and light relief.

Finally, I would also like to say a special thank you for the generous philanthropic donations that have supported my PhD and without which this body of work would not have been possible.

Abstract

An important question in cancer evolution concerns which traits make a cell likely to successfully metastasise. Through a combination of experimental evolution and computer vision a series of mathematical models have been developed throughout this thesis to investigate the individual signal processing behaviour of cancer cells during dispersal. In Chapter 2 a convolutional neural network is used to demonstrate how the morphology of individual cells can be automatically segmented within phase contrast time-lapse videos. The segmented morphologies are then used in Chapter 3 to explore the idea of signal processing mediated dispersal to reveal a density-dependent phenotype only seen in cells selected for distant site colonisation. Specifically, the model shows that the rate of morphological change is positively correlated with the speed of migration when the local cell density is high. However, when the local cell density is low the opposite relationship is displayed: the rate of morphological change decreases with an increase in migration speed. Chapter 4 then builds upon the results of Chapter 3 to develops two temporally dependent morphological model that quantify short term temporal changes in dispersal dynamics at both a population and single cell level. The temporally dependent models reveal that in fact a subset of cells in all of the experimental populations can adopt similar complex behaviour. However, the populations differ in their behavioural demography as well as the frequency at which a given behaviour is adopted through time. Finally, Chapter 5 employs a similar temporally resolved approach to investigate the interaction between the broader cancer cell population and a small subset of cancer cells known as poly-aneuploid cancer cells. In summary, this thesis harnesses the power of mature mathematical techniques to investigate novel and emergent characteristics of metastatic dispersal in a quantitative and statistically robust manner.

Table of contents

List of figures

List of tables

Glossary

Definitions

Angiogenesis	The formation of new blood vessels from pre-existing vessels
Apoptosis	Programmed cellular death
Fitness	The propensity of an organism to survive and reproduce in their current environment (Orr, 2009)
Intravasate	The movement of a cell through the well of a blood or lymph vessel and into the vessel itself
Tumorigenesis	The formation of a tumour

Mathematical notation

*	Complex conjugate
i	Unit imaginary number $\sqrt{-1}$
$\psi_{pq}(xy)$	Image function
$Z_{m,n}$	Zernike moment with order m and an integer of rotation n

Acronyms / Abbreviations

CNN	Convolutional neural network
ITH	Intratumoural heterogeneity

mAP	mean Average Precision
PACC	Poly-aneuploid cancer cell
ReLU	Rectified Linear Unit
SGD	Stochastic gradient descent
SSM	State space modelling

Chapter 1

Introduction

1.1 Cancer

Cancer is a collective term associated with over 100 forms of disease that are characterised by uncontrolled cellular growth. The onset of aggressive proliferation gives rise to a malignant mass of cells known as a tumour (Weinberg, 1996). Once formed, the tumour then develops over time through a combination of genetic and epigenetic changes. Finally, the molecular changes then culminate in advantageous phenotypic changes that are selected for producing a heterogeneous population of cells (Greaves and Maley, 2012).

The Hallmarks of Cancer

Whilst cancer encompasses a broad spectrum of diseases there are a set of phenotypic changes that are seen as essential across all cancer types, *The Hallmarks of Cancer* (Hanahan and Weinberg, 2000). Uncontrolled cellular proliferation, a loss of growth suppression, and an evasion of cellular apoptosis are 3 of the earliest hallmarks of cancer and together they permit the formation of a microscopic tumour (Polyak, 2007). However, to increase in size further, the tumour must also upgrade its corresponding infrastructure. This demand is met through the onset of sustained angiogenesis whereby new blood vessels are built so that the tumour can continue its expansion and become clinically relevant (Raica et al., 2009). Yet, as much as 90% of cancer related mortality is not solely due to the formation of a primary tumour.

Rather it is the result of cancer cells spreading around the body and forming tumours at secondary sites, a process known as metastasis (Chaffer and Weinberg, 2011).

1.1.1 Cancer evolution

The constant increase in cell population coupled with selection for the fittest variant has allowed tumour development to be understood as an instance of Darwinian natural selection, a key mechanism of evolutionary change (Greaves and Maley, 2012; Merlo et al., 2006; Nowell, 1976). Importantly, natural selection is the only process that leads to adaption on an evolutionary timescale (Barton, 2007). Hence natural selection can be used to understand how a tumour first forms, how it grows, and ultimately how it becomes resistant to therapy (Venkatesan and Swanton, 2016).

Natural selection in cancer

A tumour first forms as a result of an individual cell developing a proliferative advantage relative to its neighbours. This advantage is then selected for causing a period of rapid proliferation to ensue and a subsequent expansion in tumour size. However, during this time, random mutations also appear within the population. The vast majority of the mutations either have a negligible or negative effect causing the mutant to die out. Yet, occasionally, a beneficial mutation will arise and give the mutant a competitive advantage over its peers i.e. a higher degree of invasiveness. This advantage is then passed onto subsequent generations resulting in the mutant lineage becoming dominant within the tumour mass. A similar process can be seen when a therapeutic intervention is applied. A therapeutic intervention will kill the vast majority of cells within a tumour leaving only a small sub-population of cells to remain. However, all of the cells within the remaining sub population are therapy resistant. Thus, when the therapy ends, and the tumour regrows, the entire tumour population is now resistant to therapy (Maley and Reid, 2005; Nowell, 1976).

Criteria for selection to act

For selection to act there are certain conditions that must first be met within the population: (Lewontin, 1970):

- Phenotypic variation: Individuals within the population have different phenotypes
- Differential fitness: Different phenotypes have different rates of survival and reproduction in certain environments
- Fitness heritability: There is a correlation in fitness between parents and their offspring

If these principles hold, then a population will experience evolutionary change (Barton, 2007). In the context of cancer, phenotypic variation is present due to genetic and epigenetic variation that arises as a result of mutations, and changes in methylation. In turn, the phenotypic variation then creates a fitness differential within the tumour as seen by the heterogeneous population of cells that remain. Finally, the increase in fitness is then heritable as it is encoded into the nucleotides and methylation patterns of the DNA (Maley and Reid, 2005).

1.1.2 Cellular selection

Whilst natural selection shapes cancer progression (Greaves and Maley, 2012), selection also shapes the development of healthy functional tissue. Therefore, cancer is not characterised by the presence of natural selection, but rather the unit at which selection acts.

The unit of selection refers to the biological tier upon which selection acts: molecular, cellular, organism, or group (Lewontin, 1970). The formation of functional tissue requires successful multi-cellularity. Therefore, individual cellular selection is suppressed in favour of the organism becoming the primary unit of selection (Michod, 1999). In contrast, cancer is a loss of individual cellular control. Hence selection defaults to the cellular level where each individual cell is working to improve their own local environment and increase their own chances of survival (Rieger and Welter, 2015). This selfish behaviour is a defining feature of

cancer and it enables the disease to become extremely resilient. However, it can also have negative repercussions as seen by the poorly formed vasculature within the tumour mass.

Sustained angiogenesis

The vasculature within a tumour is developed through a period of sustained angiogenesis (Raica et al., 2009). Angiogenesis is a process commonly seen in wound healing (Tonnesen et al., 2000) and mid stage embryogenesis (Breier, 2000) to facilitate the formation of new tissue. In both cases angiogenic signalling is tightly controlled over a set period of time in which neighbouring cells are working as a collective. This results in a highly organised vasculature network being formed which can in turn provide a constant supply of resources to the tissue (Basanta and Anderson, 2017).

In contrast, the heterogeneous cell population within a tumour is not acting as a collective, each cell is acting for itself. Therefore, each cell will only invest into angiogenic signalling until an adequate level of resources have been personally received. As a result, this causes the vasculature within the tumour to form as an unordered, excessively branched, leak riddled network (Hanahan and Weinberg, 2011). The poorly designed network then leads to an erratic bloody flow within the tumour which creates both spatial and temporal heterogeneity in the distribution of resources (Gillies et al., 2018). In turn the resource heterogeneity further increases the competition between neighbouring cells.

Metastatic paradox

Intense cellular selection can explain how traits such as uncontrolled proliferation and a loss of growth suppression arise within a tumour cell population. They both increase the rate of proliferation which in turn increases the individual's fitness. Likewise, the ability to induce sustained angiogenesis improves the surrounding tumour environment which also increases the individual's fitness (Fortunato et al., 2017; Merlo et al., 2006). Yet, in contrast, the ability to metastasise appears to have no direct fitness benefit within the primary tumour. Furthermore, it may even be expected that the abundance of metastatic cells will decline at the primary tumour site as cells that are capable of metastasising will have already left the

tumour (Bernards and Weinberg, 2002). Nevertheless, metastasis occurs in nearly all types of cancer and remains a broadly incurable stage of cancer progression (Pienta et al., 2020).

1.1.3 Metastasis

Metastasis is a multi-step process whereby a cell spreads from the primary tumour and then eventually colonises at a distant site in the body. The series of stages from tumour dissemination to distant site colonisation are referred to as the metastatic cascade (Figure 1.1).

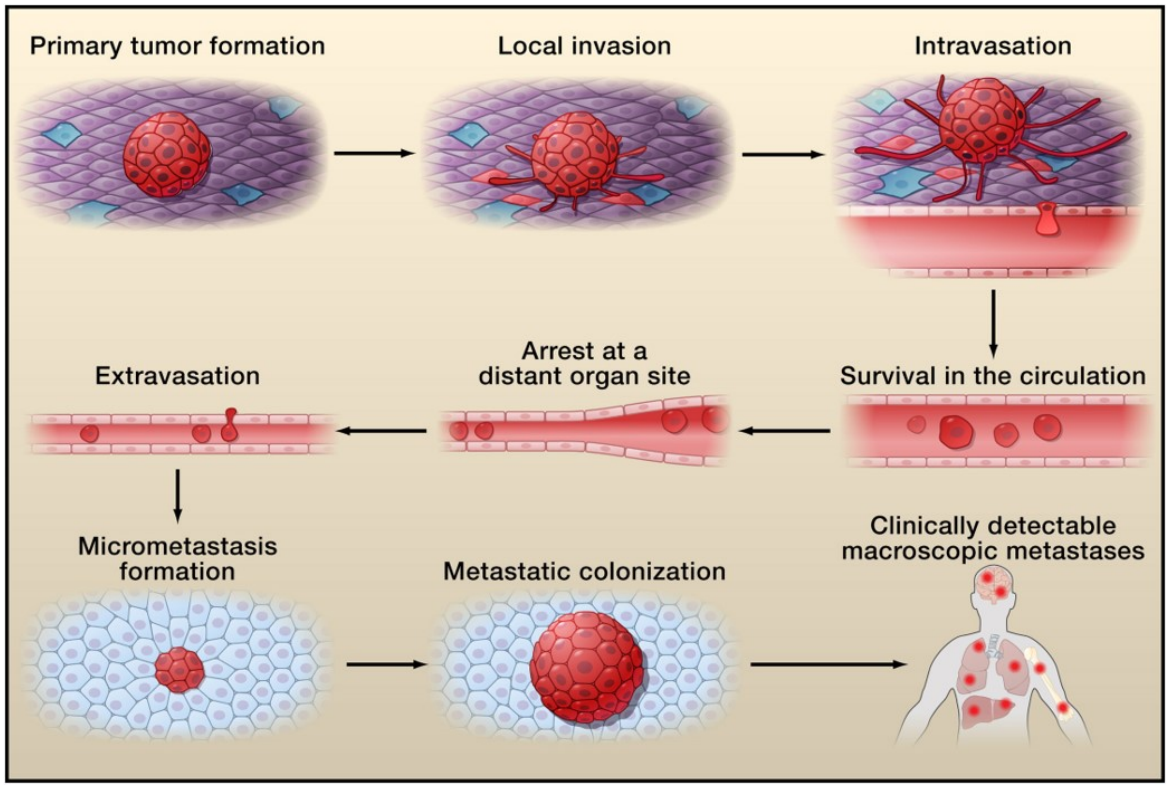


Figure 1.1: The metastatic cascade; taken from (Valastyan and Weinberg, 2011).
The metastatic cascade begins within a cancer cell first leaving the primary tumour and then invading into the local micro-environment towards a nearby blood vessel. The cell then enters into the circulatory system before being carried around the body to a distant site. Upon reaching a distant site the cell then leaves the circulatory system and forms a microscopic tumour to aid its survival. Finally, the cell then re-initiates aggressive proliferation and colonises the distant site to form a clinically relevant secondary tumour.

The cascade begins within a cell escaping from the primary tumour and migrating through the extracellular matrix towards a nearby blood vessel. Once at the blood vessel the cell then intravasates into the circulatory system before being carried around the body to a distant site. Upon reaching a distant site the cell then extravasates from the blood and invades into the foreign tissue. To actively survive within the foreign environment the cell then proliferates to form a small microscopic tumour. Finally, to colonise the distant site the cell must re-initiate aggressive proliferation resulting in the formation of a clinically relevant macroscopic tumour (Valastyan and Weinberg, 2011).

Timing metastatic progression

Whilst the stages of metastasis are well known, dating when a metastatic cell first appears within the tumour population remains unclear. Classically metastatic spread has been understood through a linear progression model where a metastatic sub-population appears late in tumour development (Weinberg, 2007). However, recent molecular evidence suggests that the formation of a metastatic sub-population might occur earlier in tumour development, referred to as the parallel progression model. The parallel progression model also suggests that metastasis may occur without the need for local tissue invasion and that a cell might intravasate straight into the tumour's vasculature (Deryugina and Kiosses, 2017; Klein, 2009).

Accurately dating when a metastatic population first appears within the primary tumour is important because it helps to estimate the degree of heterogeneity between the primary tumour and distant metastatic site. In turn, the degree of heterogeneity is a significant clinical marker because an increase in heterogeneity is seen to correlate with an increase in therapeutic resistance (Turajlic and Swanton, 2016). In regard to the two different models of metastatic progression the linear progression model is expected to have less heterogeneity between the primary and distant site tumours compared to the parallel progression model (Caswell and Swanton, 2017). However, the two models of metastatic spread are not mutually exclusive. That is, both models can occur within the same patient at different points in time (Turajlic and Swanton, 2016). Hence, when investigating different stages of the metastatic

cascade, it is important to evaluate traits at a single cell level to ensure that the heterogeneity within the metastatic population is captured.

Metastatic inefficiency

In spite of metastasis being the most-deadly stage of cancer progression, it is also considered to be the most inefficient (Chaffer and Weinberg, 2011; Chambers et al., 2002). Experimental evidence has found that only 0.02% of metastatic tumour cells injected into the circulatory system of a mouse subsequently formed a clinically relevant tumour at a distant site (Luzzi et al., 1998). Furthermore, this figure does not account for the initial dissemination from the primary tumour whereby a cell must avoid immune detection and navigate through the complex collagen rich micro-environment (Gupta and Massagué, 2006). Hence relative to the number of cells that initially disseminate, the proportion of cells that successfully colonise is expected to be considerably less.

However, the inefficiency is not believed to be uniform across the cascade. Experimental evidence suggests that process is bottle-necked by the final two stages, the ability to survive and then colonise at a distant site. After entering into the circulatory system 83% of cells managed to extravasate, but only 2% and 0.02% complete the final two stages. The remaining extravasated cells entered in to a dormant non-proliferative state known as quiescence (Luzzi et al., 1998). Cellular quiescence is a transient state (Yao, 2014) where a cell has left the cell cycle and is undetectable to the immune system (Gomis and Gawrzak, 2016).

The high rate of cellular dormancy, but lack of apoptosis, suggests that existing and thriving at a distant site are not one in the same. The ability to overpower and dominate the local population is less common compared to quietly integrating into the masses. Hence the few cells capable of distant site colonisation most likely possess a unique, but deadly, set of traits. Identifying the traits that separate the small lethal sub-population of cells from the broader metastatic population will therefore be essential in trying to constrain metastatic spread.

1.2 Dispersal theory

The fitness of an individual is measured relative to its current environment (Clobert et al., 2012). Thus, a change in environment will often cause a corresponding change in fitness. In turn, if the environment is spatially and temporally heterogeneous, then the fitness of an individual can vary extensively across multiple different locations. In such a setting ecological dispersal theory predicts that dispersal can be selected for when the gain in fitness from moving outweighs the cost of migration itself (Bowler and Benton, 2005). In short, if the fitness is higher at a different location, and the cost of moving between locations is low, then the ability to disperse can be under selection. Ecological dispersal theory therefore provides a clear explanation for the onset of metastatic spread (Amend et al., 2016).

Firstly, the underdeveloped vasculature network within the tumour means that resource levels are both spatially and temporally heterogeneous. This variability then creates a fitness differential within the tumour that causes an individual's fitness to vary in both time and space. Hence the ability to disperse can be selected for as it allows the individual to move in response to environmental changes and therefore prevents a reduction in fitness.

Interpreting metastatic spread in the context of dispersal provides a powerful framework to investigate the different environment pressures that can select for a dispersal phenotype as well as the corresponding types of dispersal i.e. long vs short range (Bonte et al., 2012). The following section discusses different ecological pressures that can select for a dispersal phenotype and how they relate to the evolution of metastasis. The section then concludes by highlighting the importance of accurate signal processing during dispersal and its application in metastatic spread.

1.2.1 Resource availability

If resource availability is constant, regardless of the severity, then selection favours a specialised phenotype to fill the evolutionary niche (Futuyma and Moreno, 1988). However, in many environments, as within a tumour, resource availability can fluctuate both spatially

and temporally. In turn, this variability can select for a phenotype that is equally as flexible, phenotypic plasticity (Gillies et al., 2018).

Phenotypic plasticity

Phenotypic plasticity involves a genotype producing different phenotypes when exposed to different environmental conditions (King and Hadfield, 2019). Plasticity therefore allows a genotype to be more tolerant to changes in the environment and thus have a higher fitness across multiple environments (Ghalambor et al., 2007). If the variability is temporally predictable then this flexibility can allow for maximum reproduction in times of resource prosperity (Wang et al., 2015) coupled with dormancy during times of resource austerity.

In contrast, if resource availability is temporally stochastic then plasticity can arise via dispersal (Bowler and Benton, 2005). Resource driven dispersal occurs throughout ecology and has been shown experimentally in cancer cell populations to select for increased motility and dispersal like behaviour (Chen et al., 2011; Taylor et al., 2017). Furthermore, if an environment is also spatially heterogeneous, such as in a tumour, then either a conditional or unconditional dispersal strategy can be adopted.

Unconditional dispersal

An unconditional dispersal strategy means that the rate of dispersal, the number of migrants entering and leaving an area, is constant irrespective of the environmental conditions (McPeek and Holt, 1992). Unconditional dispersal can therefore be seen as a form of bet-hedging where an individual aims to lower their own fitness variance (Villa Martín et al., 2019). This means that in an optimal environment there is a reduction in the maximum level of fitness so that in a sub-optimal environment there is a corresponding increase in fitness (Rees et al., 2009).

An example of bet-hedging can be seen in the form of seed banks whereby a proportion of seeds emerge at different points in time. This means that the maximal level of germination is never achieved, and neither is the maximal level of fitness, as certain seedlings will emerge in sub-optimal conditions. However, if all of the seedlings are destroyed at a given point in

time e.g by an extreme weather event. Then the plant has a second chance at producing viable descendants from the subsequent waves of seedlings (Fan et al., 2018). An unconditional dispersal strategy also provides a similar reduction in fitness variance. The continual dispersal means that an individual can end up leaving a prosperous region before all of the resources have been exploited, thus incurring a drop in maximum fitness. However, it is also means that the migrant will move straight through an unfavourable region and thus only be exposed to the adverse conditions temporarily (Gillies et al., 2018).

Conditional dispersal

In contrast, conditional dispersal is more complex. The decision to disperse is dependent on one or many environmental cues being satisfied e.g. local resource availability drops below a minimum threshold. In turn, this means that under certain environmental conditions multiple dispersal strategies can evolve and exist at an evolutionary equilibrium (McPeek and Holt, 1992). In the context of cancer, this is problematic as it means that to prevent dispersal as a whole you need to block each individual strategy (Katt et al., 2018). Yet, to block each strategy requires a detailed understanding of the individual phenotypes within the population as well as how they interact. In a complex disease such as cancer this is extremely difficult as phenotypic variation within a population is often very diverse. Also, individual cells are known to transiently change their phenotypic traits (Section 1.3). Hence containing a migratory population relies upon finding a trait that is common among the different dispersal strategies.

1.2.2 Population density

Whilst dispersal can be selected for due to a change in resource supply it can also be selected for by a change in demand (Matthysen, 2005). This change can occur through a local increase in population density, or through an increase in one region coupled with a decrease in another. In either case a fitness differential is caused regardless of whether the environment is constant or variable (Hamilton and May, 1977).

Density dependent dispersal in early tumour development

Dispersal driven by a change in population density tends to be more prevalent in smaller populations (Bowler and Benton, 2005). This is due to a local increase in population size having a larger relative impact and therefore the corresponding increase in competition is greater. As a result, this might provide an explanation for the onset of metastasis early in tumour development, as seen in the parallel progression model. Initially the tumour is small both in regard to its spatial dimensions and its population. Therefore, a burst in proliferation locally would have a larger impact on the topology of the tumour, and thus the competition within.

The lack of local invasion in a parallel progression model might also be explained by density driven dispersal. If the increase in population occurs locally, and therefore the reduction in fitness has affected those locally, an equal increase in fitness can be gained by dispersing locally. In contrast, dispersal that is driven by a decrease in resource availability tends to be longer as the migrant needs to find a new pasture (Clobert et al., 2012).

1.2.3 Kin selection

Dispersal can be selected for by a change in environmental conditions, but likewise dispersal can also change the environment in its wake. As a migrant leaves the native site the remaining population benefit from a reduction in competition. This reduction may be coincidental. However, it may also be an active process to improve the fitness of a close relative, kin selection.

Kin selection occurs when selection acts on traits to improve the fitness of a relative at the expense of the own individual's fitness (Smith, 1964). Hamilton's rule states that kin selection should act when the benefit to a recipient, accounting for the level of relatedness r (Wright, 1922), outweighs the cost to an actor. Formally, Hamilton's rule can be written as $rB > C$, where r is the level of relatedness between the recipient and the actor, B is the gain in fitness to the recipient and C is the fitness cost to the actor (Hamilton, 1964). Therefore, as

the level of relatedness diverges the probability of kin selection decreases. In a highly related population such as cancer kin selection may therefore provide a strong selective pressure.

Kin selection dispersal range

Dependent on the rate of genetic divergence kin selection could potentially drive selection for dispersal in both the linear and parallel progression models of metastasis. Firstly, there needs to have been enough time for the level of relatedness between individual cells to decrease. The level of relatedness r within the cancer cell population is ≈ 1 at the onset of tumour development due to the clonal nature of cancer. As a result, kin selection may not be viable as the fitness increase locally would be outweighed by the individual cost of migration, and the decrease in fitness for the clonal relatives that receive the migrant. Hence in the parallel progression model this would explain the delay in the onset of a dispersal phenotype (Turajlic and Swanton, 2016).

In contrast, in the linear progression model, kin selection may explain the onset of a longer range. At a global level, the tumour is genetically diverse. However, as the tumour grows in size the cells that are spatially close to one another are more likely to be genetically similar. Hence increasing the dispersal distance increases the probability that the area that receives the migrant is genetically divergent from the migrant itself, albeit the cost and risk to the migrant is higher. However, elucidating whether a change in kin structure is driving selection for dispersal is problematic. Firstly, kin structure is typically intertwined with other selective pressures such as resource quality and therefore pinpointing its exact affect can be extremely challenging (Bowler and Benton, 2005).

1.2.4 Signal processing

To summarise, a variety of different ecological pressures can select for the ability to disperse. Likewise, the same ecological pressures can select for a myriad of different dispersal behaviours and habitats for subsequent colonisation (Clobert et al., 2012). Yet, common to all aspects of dispersal is the need for an individual to be able to receive and respond to changes

within their environment, a process known as individual signal processing (Clobert et al., 2009).

Signal processing in cancer

The action of receiving, evaluating, and responding to both public and private signals can also be seen during metastasis:

1. Emigration

   Firstly, aggressive cellular proliferation is known to create a hypoxic and highly acidic tumour core (Amend and Pienta, 2015). In turn, the deterioration in environmental conditions can increase the fitness of certain tumour sub-populations. However, the reduction in oxygen availability has also been proposed as a leading cause of metastatic dispersal (Amend et al., 2016). Hence whether an indiviudal cell stays or goes depends on the cell detecting the current local oxygen levels.

2. Migration

   Secondly, a cell is guided by environment signals to shape its migratory behaviour (Section 1.3). The direction in which a cell migrates is in response to both chemotactic and durotactic gradients (Alberts et al., 2008). The gradients can be geographically fixed, as in the case of a capillary, or they can change location dynamically as seen during cellular streaming (Section 1.3). Likewise, the type of migration that a cell adopts i.e. ameoboid or mesenchymal is dependent on the signals that are received from the substrate that a cell is moving along (Friedl and Wolf, 2010).

3. Immigration

   Finally, certain cancer types are known to colonise at particular organs, the "seed and soil" hypothesis (Paget, 1889). Similarly, once at the distant site, individual cells are known to reside in a state of prolonged cellular dormancy before re-initiating aggressive proliferation (Giancotti, 2013). Hence both aspects rely upon a cell responding to

their current environment conditions and then responding accordingly. In other words, processing an incoming signal.

However, in contrast to an organism, individual signal processing in cancer is a recent life history event. Historic mechanisms exist within the cell, but they have been broadly repressed so that multi-cellularity can flourish. An example of which can be seen during the onset of sustained angiogenesis. Hence the re-activation of specific intra-cellular pathways present an opportunity for variation to arise within the population, and thus a potential substrate for selection to act. As a result, individual signal processing maybe a key determinate in metastatic success and a possible means by which to identify cells with increased metastatic potential.

Morphological proxy

To explore the role of signal processing in metastasis requires the quantification of signal exchanges on a cellular level. Yet, due to the number of different environment signals that are present, and the corresponding number of individual receptors on the surface of a cell, it is not tractable to sequentially test the effect of each individual protein pathway. An alternative approach, that encompasses the entire spectrum of possible signals, is to use morphology as a proxy (Tweedy et al., 2013). A change in morphology then represents the transmission or receipt of a signal and the magnitude of morphological change correlates with the signal intensity. This approach also benefits from being able to measure morphology in situ. Thus, the signal emission from dynamic factors such as the location of a neighbouring cell can also be captured. The next step is to understand how an individual cell responds to different environmental signals and how those signals can influence its phenotypic behaviour.

1.3 Cell migration

Cancer cell migration is a complex process that involves the coordinated activity of hundreds of proteins in order to generate cell polarity, actin filament polymerisation, focal adhesion

turnover, and cellular traction (Butler et al., 2019). A detailed knowledge of the stages before, during, and after a cell migrates is therefore essential to understand how cancers metastasise.

1.3.1 Stages of migration

Gaining a motile phenotype

Firstly, to facilitate migration, most cancer cells are required to lose cell-cell adhesion and epithelial polarity whilst gaining motility and invasiveness, a process that is referred to as *Epithelial - Mesenchymal Transition* (EMT) (Tam and Weinberg, 2013). EMT is a reversible phenotypic change normally associated with embryogenesis and wound healing that depends on a diverse network of epigenetic mechanisms (Kalluri and Weinberg, 2009). After having undergone EMT a cell is then able to squeeze through gaps within the tightly packed ECM and is only further limited by the size of its nucleus (Wolf et al., 2013). However, EMT is not binary. Experimental evidence indicates the EMT is a multi-step process and as such the degree of EMT varies across the population (Grigore et al., 2016). As a result, the corresponding degree of invasiveness and motility also varies across the cancer cell population.

Movement mechanics

Next, once a cell has become motile its individual movement is initiated in response to external signals that are received by receptor proteins on the cell membrane (Kim et al., 2011). The signal is then transported via a complex network of pathways to the interior of the cell (Alberts et al., 2008). Next, the cell proceeds to rearrange its cytoskeleton and form a pseudopodia. A pseudopodia is an actin based structure that forms at the leading edge of the cell. Once the pseudopodia has formed it adheres to the substrate and provides a direction for the cell to move in (Lauffenburger and Horwitz, 1996; Olson and Sahai, 2008). The cell then physically moves by detaching from the substrate at its rear. The subsequent contractile forces then drive the cell forward onto the newly created pseudopodia and then the process repeats (Alberts et al., 2008; Ananthakrishnan and Ehrlicher, 2007; Cameron et al., 2000).

Cell-Cell adhesion	Cell-ECM adhesion	Cellular contraction	Migration mode
High	High	Moderate	Collective strand
High	Moderate	High	Collective sheet
Low	High	High	Mesenchymal
Low	Moderate	Moderate	Amoeboid pseudopodal
Low	Low	Low	Amoeboid blebby

Table 1.1: Modes of Cell Migration:
The resultant modes of migration from different combinations of cell-cell adhesion, cell-ECM adhesion and cellular contraction (Friedl and Wolf, 2010).

Environmental repercussions of migration

Finally, the aftermath of a migratory cell population can be seen by the topological changes that occur within the surrounding micro-environment (Fang et al., 2014). During tumour dissemination, and the subsequent invasion at a distant site, a cell is required to navigate through the stroma that separates the circulatory system from the basement membrane (Clark and Vignjevic, 2015). The stroma, among other things, provides a scaffold for binding together tissue and cells through a plethora of proteins termed the extra cellular matrix (ECM), of which collagen is the most abundant constituent (Bremnes et al., 2011; Kohn et al., 2015).

A change in collagen composition is one of the earliest clinical markers of cancer progression. Initially, the collagen density increases around the tumour causing the surrounding matrix to stiffen. The increased stiffness then allows the tumour to continue its expansion by displacing the host tissue (Gkretsi and Stylianopoulos, 2018). After the collagen density has increased, it is then followed by a straighten of fibres perpendicular to the tumour boundary (Figure 1.2) (Provenzano et al., 2008). The straightened fibres are then bundled together to form "highways" that cells are able to exploit and thus move through the ECM in a more directed fashion (Wershof et al., 2019). The change in collagen composition demonstrates the symbiotic relationship between migratory cells and the surrounding micro-environment that continues throughout metastatic dispersal (Yuan, 2016).

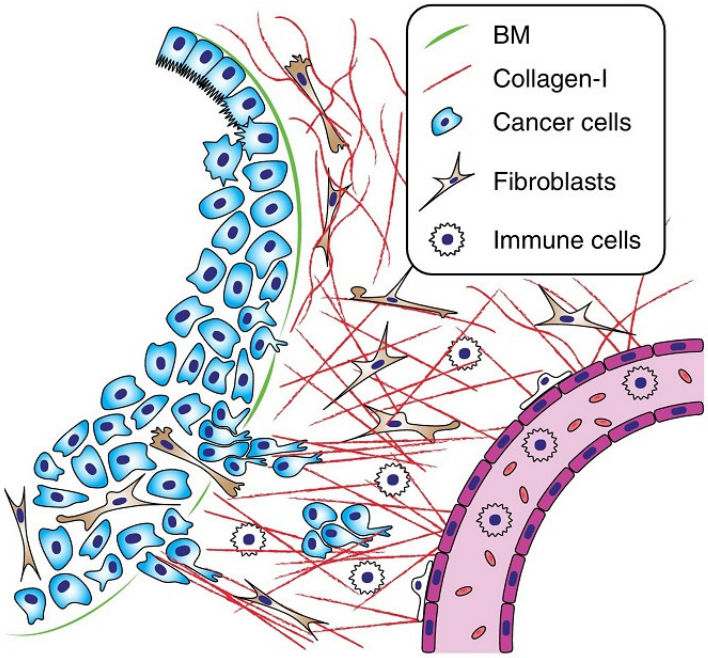


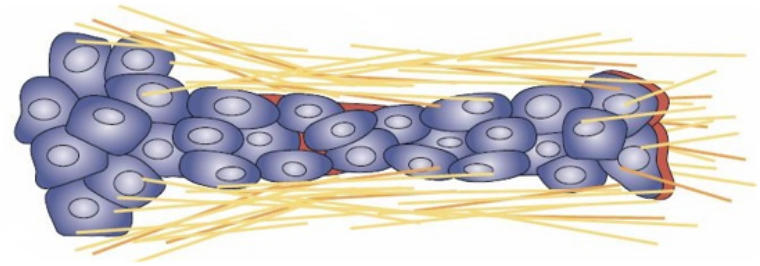
Figure 1.2: Cancer cells leaving the primary tumour; adapted from (Clark and Vignjevic, 2015). A graphical image of cancer cells escaping from the primary tumour (left of the figure), migrating through the basement membrane, and into the stroma. Once in the stroma the cells then migrate towards a blood vessel in the bottom right. The point of escape has an increased immune response and a straightening of the local collagen fibres. In addition, the cells at the top of the figure are uniform and are adhered to one another. In contrast, the cells at the point of escape are more irregular and elongated with a lack of cell to cell adhesion, characteristic features of cells that have undergone Epithelial - Mesenchymal Transition.

1.3.2 Modes of cell migration

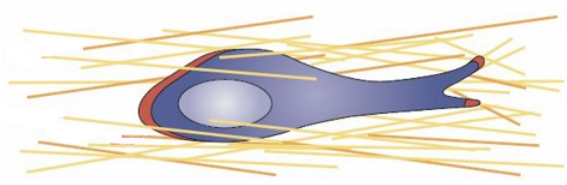
During migration cells adopt a wide variety of different migratory modes. The mode of migration depends on the degree of cellular adhesion, to both other cells and the ECM, as well as the degree of cellular contraction (Table 1.1) (Friedl and Wolf, 2010). Broadly there are 3 types of migration although multiple cells can also migrate in groups known as multi-cellular streams (Friedl and Wolf, 2003; Paul et al., 2016) (Figure 1.3):

- Collective migration

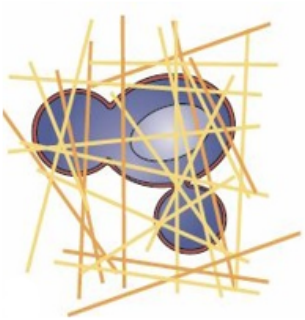
  Collective cell migration occurs when cell-cell adhesion is still present and is often observed at the edge of the tumour boundary (Friedl et al., 2012) (Figure 1.3a). Collec-



(a) Collective strand migration



(b) Single cell mesenchymal migration



(c) Single cell amoeboid migration

Figure 1.3: Modes of cancer cell migration; adapted from (Friedl and Wolf, 2003). Broadly cancer cells use 3 types of migration: a) Collective migration where cell-cell adhesion remains, b) Single cell mesenchymal migration where cells have lost cell-cell adhesion but have a high degree of adhesion to the substrate, c) Single cell amoeboid migration where cells have lost cell-cell adhesion and have moderate to low adhesion to the substrate.

tive cell migration can occur as either a linear strand with a single leading cell or as a sheet with multiple leading cells (Clark and Vignjevic, 2015).

- Single cell migration

  If cell-cell adhesion is no longer present then the mode of migration is predominately dependent on the degree of cellular adhesion to the matrix. A high degree of cell-ECM adhesion produces a mesenchymal phenotype that is characterised by a *"spindle-shaped"* morphology with well defined protrusions know as pseudopodia (Figure 1.3b). In contrast, amoeboid migration occurs when the degree of cell-ECM adhesion is low and is identifiable by a more rounded morphology with a poorly defined leading edge (Pandya et al., 2017) (Figure 1.3c)

- Multi-cellular streaming

  Whilst amoeboid and mesenchymal migration are types of single cell migration they can also occur within higher level group structures known as multi-cellular streams (Friedl et al., 2012). In contrast to collective migration where the cells are physically joined, the cells within a stream remain independent (Gaggioli et al., 2007). This allows the stream to quickly change its formation and adapt to the environment in which it is moving through. As a result, multi-cellular streaming is one of the fastest and most directed modes of migration with an average cell speed of $\approx$ 1–2$\mu m/min$. In comparison, solitary migration is approximately an order of magnitude slower and collective migration can be up too two orders of magnitude slower, 0.2–0.4$\mu m/min$ and 0.01–0.05$\mu m/min$ respectively (Clark and Vignjevic, 2015).

Benefits of group dispersal

The process of disseminating from a primary tumour and surviving at a distant site are both functionally similar. They involve a cell migrating through the stroma and reaching a target location, either a blood vessel or an area suitable for colonisation (Figure 1.1). Nevertheless, there is a large difference in their relative success (Luzzi et al., 1998). However, due to the lack of cellular apoptosis, the critical difference is assumed to be essential for metastatic progression but not critical for cellular survival.

One possible explanation might be the number of cells that are present within the local vicinity. During dissemination there is a high density of cancer cells migrating locally and the route has been down-trodden previously by earlier waves of migrants. In contrast, when arriving at a distant site the probability that another malignant cell has been in the area is approximately 0. Hence cancer cells may rely upon a community approach to successfully survive and then colonise at a distant site. This may also explain why the presence of circulating cellular clumps is indicative of a worse prognosis (Murlidhar et al., 2017). That is, when a clump arrives at a distant site the group structure is already present and therefore any cooperative strategies can be enacted straight away. If true, and group structure is critical for metastatic success, then accurate neighbour communication will be an essential

prerequisite. As a result, this would reaffirm the importance of individual signal processing during metastatic dispersal.

1.3.3 Migratory dynamics

To determine whether group level cooperation is critical for metastatic success relies upon having a thorough knowledge of the individual phenotypes that exist within the population. In the context of cancer, phenotypic information is commonly gleamed by measuring traits from fixed cell images. In turn, the measured traits are then used to summarise the phenotype of a cell and categorise its behaviour (Table 1.1) (Gordonov et al., 2016; Meijering et al., 2012).

Cell state

Whilst this is approach can be informative it also assumes that the phenotype of the cell remains constant over a short period of time i.e. within a single generation. Yet, in reality, a cell's phenotype can, and often does, vary over shorter time periods by spanning multiple different *states* (Adler and Sánchez Alvarado, 2015). The state referees to the individual realisation of a phenotype. For example, EMT is a transition from an epithelial to mesenchymal phenotype. Yet, in between the two defined phenotypes there is also an array of hybrid states that a cell can transiently express (Figure 1.4) (Lambert et al., 2017). Capturing this within cell variation is important when investigating higher level behaviours as cooperative groups often have a precise population structure. Hence small deviations at an individual level can manifest into large shifts in population level dynamics. Also, defining a phenotype as a collection of states provides an important link between the multiple levels of variation that can exist within a population at different timescales i.e. phenotypic plasticity on an evolutionary timescale vs state changes on a generational timescale.

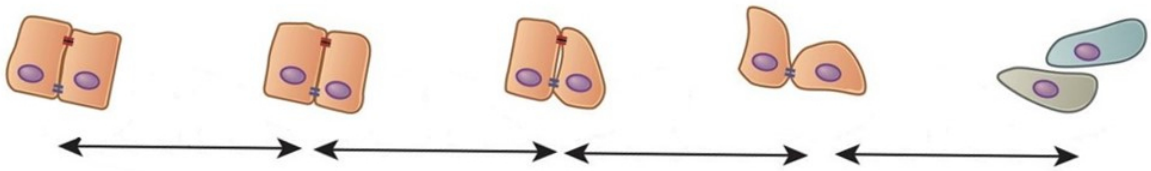


Figure 1.4: Hybrid EMT states; adapted from (Lambert et al., 2017).
The epithelial mesenchymal transition (EMT) of two cells from an epithelial phenotype, left of the figure, to a mesenchymal phenotype, right of the figure. The 3 intermediate figures demonstrate the hybrid states that a cell can adopt between the two fixed phenotypes.

Geographical interpretation of cell states

Similar to Waddington's epigenetic landscape (Goldberg et al., 2007) the relationship between phenotypes and states can also be represented in a geographic context. The different possible phenotypes are represented as countries joined together on a map. The possible states within each phenotype are then analogous to the different locations within a country that a person can live. Feasibly a person can live in any part of the country. Yet, in practice, certain areas are more common i.e. cities, towns etc. Likewise, all states are possible within a given phenotype, but certain states are more stable and therefore more common. Furthermore, moving to a new city within a country tends to be more prevalent and is often a faster process than moving between separate countries. Similarly, a change in cell state occurs over a shorter time period relative to a change in phenotype. Hence developing a thorough understanding of the state landscape will be useful to predict which states will flourish within certain temporal environments.

Cell morphology

A marker of cell state, and an emergent property of cell - environment interaction, is captured by a cell's morphology (Prasad and Alizadeh, 2018; Rangamani et al., 2013; Wu et al., 2020). Historically qualitative measures of morphology in fixed images have been used to differentiate between different cellular phenotypes. However, more recently, a quantitative approach has been adopted allowing for a greater degree of precision and more subtle changes in phenotype to be detected (Alizadeh et al., 2016; Lyons et al., 2016).

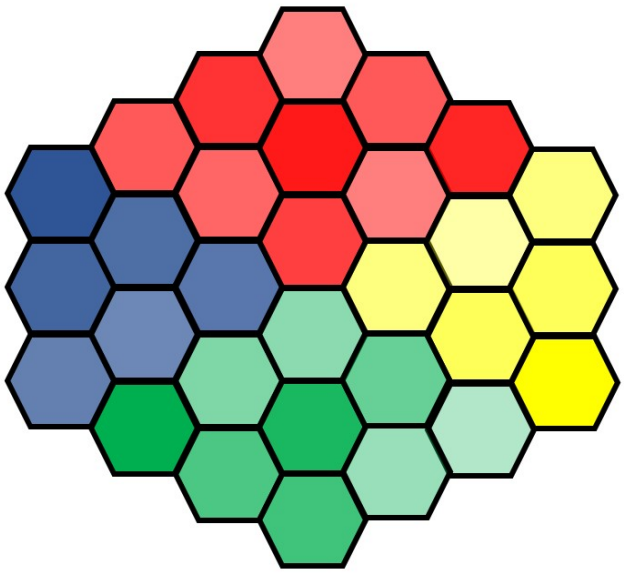


Figure 1.5: Geographical interpretation of cellular states.
The 4 main colours (red, yellow, green, and blue) represent the 4 phenotypes that a cell can exist in. The shaded hexagons within each main colour then represents the hybrid states within each phenotype. Generally, the states within a given phenotype are closer to one another and thus it is easier for a cell to move between states than between phenotypes. Yet importantly all states are still accessible from everywhere within space.

Measuring morphology in a quantitative manner also means that the variation between phenotypes can be represented as a continuum rather than discrete entities. This then allows hybrid cellular states to be investigated as well as fixed phenotypes. Likewise, measuring morphology through time also allows the transition between different states to be captured. Hence additional factors that affect cellular morphology, such as speed of migration (Lauffenburger and Horwitz, 1996; Olson and Sahai, 2008) and orientation (Tweedy et al., 2013), can also be accounted for in a dynamic setting but not necessarily through the use of fixed images.

1.4 Measuring cancer evolution

1.4.1 Retrospective analysis

Evolutionary processes, especially at a whole organism level, are not usually observable. Therefore, present day information is commonly supplemented by fossils, along with advanced genetic and statistical methods, to infer an image of the past.

The same approach has dominated the field of cancer evolution where molecular information is harvested from tumour biopsies (Graham and Sottoriva, 2017). The molecular information has then been used to investigate multiple aspects of tumour evolution such as the degree of intratumoural heterogeneity (Andor et al., 2016). In turn, the degree of intratumoural heterogeneity has then been used to try and predict whether a tumour will become invasive (Maley et al., 2006) or resistant to therapy (Morris et al., 2016). Likewise, inferring the evolutionary trajectory that leads to the increased tumour diversity has been essential in detecting large macro-evolutionary bursts that often proceed major oncogenic events (Sottoriva et al., 2015).

Limitations of retrospective analysis in cancer

However, there are also limitations with adopting a retrospective approach to understand the evolution of cancer. Firstly, fossils are not present within a tumour. Unlike most evolutionary processes cancer evolution does not possess a fossil record. This means that evolutionary "dead-ends" cannot be detected and thus all inference is based upon the lineages that survived. As a result, it is often not possible to fully understand the ecological pressures that shape the subsequent evolutionary trajectory of cancer progression.

Secondly comparative studies have focused primarily on the genetic information within a tumour rather than also considering the traits at a phenotypic level. This can be problematic as genetic changes do not necessarily manifest into phenotypic changes. Likewise, multiple different genetic changes can result in the same phenotypic change. Hence developing a thorough understanding of the genotype - to - phenotype mapping in cancer is essential to fully understand the evolutionary routes to therapy resistance (Graham and Sottoriva, 2017).

Also, due to the dynamic nature of cancer it is often difficult to detect certain transient or complex behaviours solely through a genetic lens.

1.4.2 Experimental evolution

Alternatively, some evolutionary processes can be observed directly under controlled conditions through a technique known as experimental evolution. Experimental evolution is often found in microbial biology and involves applying experimental selective pressures to replicate populations with a fast generation time. In turn, the evolutionary trajectories of the population can then be observed in real-time (Elena and Lenski, 2003). Furthermore, experimental evolution also allows the operator to precisely control the surrounding ecology both in terms of the structure as well as the resource availability. Hence specific pressures can be applied and then direct evolutionary hypothesis can be tested.

Another key advantage of experimental evolution is the ability to transfer populations to new environments and determine how their fitness changes under different selective pressures (Kawecki et al., 2012). This aspect of experimental evolution is extremely beneficial when exploring the evolutionary dynamics of metastasis, a process that inherently involves a change in environment. Likewise, it also means that experiments can be designed so that individual cells can be measured at multiple points in time. Hence removing the issue of genotype - to - phenotype mapping and capturing the phenotypic variation across multiple different timescales (Gerrish and Sniegowski, 2012).

Drawbacks of experimental approach

Nevertheless, as with any experimental system, there are some drawbacks. One of the major drawbacks with experimental evolution is the lack of transferability to complex real-world scenarios (Buckling et al., 2009). This is especially relevant when trying to study the evolutionary dynamics of a cancer, a process that is inherently very complex and dependent on the surrounding environment. Also, rather paradoxically, it is extremely challenging to include variation between experiments similar to the variation that is seen between patients in a clinical setting. As a result, the number of evolutionary trajectories

maybe underestimated compared with clinical data. Nevertheless, experimental evolution in cancer offers an exciting approach to investigate and understand what is meant by a metastatic cancer phenotype (Sprouffske et al., 2012; Taylor et al., 2013).

1.5 Thesis plan

The aim of this thesis is to explore the evolution of individual signal processing in metastatic dispersal. Through the use of experimental evolution and time-lapse microscopy the dispersal behaviour of experimentally selected populations of cancer cells have been captured. Then, through the use of morphology as a proxy, the individual signal processing dynamics have been modelled mathematically in response to dynamical features such as the cell speed and distance to the nearest neighbouring cell. As a result, the mathematical models have highlighted new and interesting biological phenomena that may have direct implications for understanding the route to metastatic success.

The next chapter, Chapter 2, details an array of computational methods that can be used to extract quantitative measures of cell migration from time-lapse videos. The chapter explains how a convolutional neural network can be used to automatically segment the morphology of individual cells. The chapter then compares the different factors that can be optimised in a convolutional neural network to improve the segmentation performance. Finally, the chapter concludes by highlighting a few key aspects that need to be considered when deploying a convolutional neural network in time-lapse videos.

Chapter 3 then focuses on the morphological dynamics at a population level. The chapter begins by detailing how the segmented morphologies in Chapter 2 can be quantified uniquely. A linear model is then built to evaluate the different morphological dynamics in response to both the speed of migration and the distance to the nearest neighbouring cell. The chapter then concludes by explaining the model results with respect to the environment selective pressures that were applied to each population.

The penultimate results chapter, Chapter 4, then builds upon the model in Chapter 3 but at a single cell level. The chapter utilises the individual time series that is recorded for each

cell during migration to investigate how the average morphological behaviour of a population varies over time. The chapter then evaluates the morphological behaviour of each individual cell to quantify the degree of phenotypic heterogeneity within each experimental population.

Finally, Chapter 5 builds on the two previous chapters to investigate the effect of transient interactions with neighbouring cells. More specifically, the chapter looks at the role of a poly-aneuploid cancer cells (PACCs) as possible conductors of metastatic dispersal. The chapter utilises a similar cell specific modelling approach as Chapter 4 but with a greater focus on the temporal behaviour of each cell. The thesis then concludes with a general discussion to highlight the key findings and to suggest a few specific areas of cancer evolution in which a similar modelling approach maybe beneficial.

Chapter 2

Convolutional neural networks as a method for automated cell tracking

2.1 Cell tracking

To quantify the phenotypic behaviour of individual cancer cells during metastatic dispersal requires imaging the migration of live cells. Phase contrast microscopy is the go-to method in live cell imaging owing to its ability to illuminate the cell, and certain internal components (Zernike, 1942). If consecutive images are taken and then coalesced together a phase contrast time-lapse video can be formed capturing the position of individual cells through time and space.

The first task when working with time-lapse data is to translate the information encoded within an image, such as the location of a cell, into a quantitative value that can be used for down-stream analysis. This pre-processing stage is known as cell tracking. The process of cell tracking can then be further broken down into two phases: segmentation and object linking. Segmentation is the first phase of cell tracking and it involves identifying which pixels are associated with each cell in a frame (Tay et al., 2010). Metrics such as the speed or direction can be obtained by segmenting a single pixel for each cell (Meijering et al., 2009). Yet, more complex metrics related to the morphology of a cell require the entire contour to be segmented. As a result, thousands of pixels need to be identified for each cell (Moen

et al., 2019). Object linking is then the second phase of cell tracking whereby the same cell is linked between consecutive frames to reconstruct its migration trajectory. If multiple cells are linked between frames, then the migration behaviours can also be interrogated at a population level as well as the single cell level (Svensson et al., 2018).

Cell tracking is a process that can be conducted either manually or automatically. Manual cell tracking is performed by an operator tracking each individual cell through time. It is a highly accessible method, but it is also slow and extremely time expensive as human input is required throughout. It also renders more complex morphological analysis intractable. Furthermore, manual tracking has a high degree of variability between operators as well as within the same operator (Huth et al., 2010). Alternatively, automated cell tracking uses a computer algorithm to track multiple cells through consecutive images. This approach is typically faster and less time expensive as a computer can track the cells without constant input from the operator.

Challenges of automated segmentation

Although appealing, automated cell tracking in phase contrast images is also extremely challenging. One of the main challenges to overcome is the low contrast between the cell and background. This is especially problematic when investigating features related to the morphology of a cell as the contrast between the cytoplasm of the cell and background is even less clear (Vicar et al., 2019). To try and mitigate this issue, and accentuate the contrast, experimental adjustments are often made through the use of fluorescent tags. Although effective, such modifications are not without limitations. Firstly, the tags tend to decrease in emission strength over time which may bias which cells are tracked. Secondly, the tags can act as another round of selection due to their toxicity (Liu et al., 1999). The combination of these two factors render fluorescent tags unsuitable for long term experimental evolution studies. Nevertheless, the following section highlights a variety of computational approaches that can used to improve the segmentation performance without any experimental modifications (Baltissen et al., 2018; Caicedo et al., 2017; Van Valen et al., 2016).

Object linking

If perfect segmentation is achieved, such that all of the cells in every frame have been identified, object linking performance is dependent on the degree of cellular displacement between frames. If cellular displacement is small, and thus the similarity between frame is high, all of the cells can be reliably linked. In contrast, if the displacement is high, multiple cells in frame t can be joined to a given cell in frame $t+1$ and thus the one-to-one matching necessary for accurate tracking is lost. This issue can however be mitigated by ensuring that the frame rate of the time-lapse video, how often an image is taken, is kept sufficiently high (Masuzzo et al., 2016). This therefore keeps the displacement between frames low regardless of the cell speed. Furthermore, a high frame rate also decreases the error between the observed migration track of the cell and the true migration track of the cell, thus increasing the accuracy of the recorded trajectory.

2.1.1 Automated segmentation

Automated segmentation has two stages; feature computation and feature selection (Berg et al., 2019). Feature computation captures the information that is encoded in an image and translates it into a numerical value i.e. the colour and intensity of a pixel or the length of an object. Feature selection then builds a model from the extracted features that can be used to segment cells in future unseen images. The parameters for each feature in the model are estimated dependent on their discriminatory power, the higher the power the bigger the weighting (Erickson et al., 2017). For example, the long continuous curve of pixels that represent the tail of a sperm cell are more informative then the small group of pixels that represent the head. As such, the tail would be given a greater weighting when classifying the object.

In cell tracking automated segmentation can be broadly divided into 3 groups dependent on the level of explicit prior knowledge from the operator:

1. ***Model based:***

Model based segmentation uses classical techniques such as thresholding and edge detection (Canny, 1986; Roeder et al., 2012; Sezgin and Sankur, 2004) to compute features. The computed features are often based upon characteristics that are deemed to be important by the operator such as the area or pixel colour of the object. Likewise, prior knowledge is used in the feature selection stage whereby a threshold is set to discriminate between groups i.e. the maximum length of a cell (Nixon and Aguado, 2012). If accurate prior information is known, that also possesses a high discriminatory power i.e the colour of a fluorescent tag, then model based segmentation can be a quick and effective method with a short lead in time. However, model based segmentation can struggle when there is large variation in the object features or the discriminatory features of an object are not already know. Unfortunately, both of these aspects are common place in phase contrast images of cancer cells.

2. ***Machine learning with features****:*

   Machine learning with features utilises the same classical thresholding techniques for the feature computation stage as the model based approach. However, the feature selection stage uses a machine learning algorithm to build a data drive model rather than relying on the operator (Sommer et al., 2011). The model is optimised through an iterative training process that uses a set of manually annotated images to search for the best solution in the data (Kotsiantis et al., 2006). The resultant model often contains non-linear unintuitive combinations of features that possess a high discriminatory power (Goodfellow Ian et al., 2016), many of which can be over looked in a model based approach.

   However, the performance of the model is also extremely dependent on the quality of the training data. The training data must be representative of the population and the manual annotations need to be of a high standard. If not, the model will fail when it is deployed on future unseen data. Furthermore, machine learning with features is still very dependent on the quality of the feature selection. This can prove to be problematic

when the cell density is high as accurate predefined features cannot be extracted from the image.

3. ***Convolutional neural networks**:*

   Convolutional neural networks (CNNs) extend the machine learning with features framework further by using data to drive the feature computation as well as the feature selection stage. CNNs compute features by first deconstructing an image into multiple levels of abstraction. The abstractions are then used to look for low level features such as edges and curves before later looking for higher order features such as shapes (Krizhevsky et al., 2012; Lawrence et al., 1997). Also, in contrast to the two previous segmentation methods, the feature computation and feature selection stages aren't necessarily disjoint. A CNN can alter the features that are computed dynamically based upon the results of the feature selection (Goodfellow Ian et al., 2016). Recent hardware improvements have meant that CNNs have been able to produce extremely high levels of accuracy in many computer vision based tasks with medical imaging, such as cell segmentation, being a prominent field (Kayalibay et al., 2017). The process by which a CNN segments an image is described in detail in the following section.

   The virtue of a data driven approach can also be the downfall of a CNN. Firstly, training a CNN requires manually annotated data, similar to machine learning with features. Yet the quantity of data that is needed is much larger. This is a major issue when working with CNNs as the generation of such data is often expensive and time consuming. Secondly it is extremely difficult to decipher which features are being computed by a CNN. As a result, if the images used during training are not representative of the population then the results from a CNN can be widely inaccurate when deployed on unseen images (Moen et al., 2019).

The focus of this thesis centres around the use of morphology as a proxy for signal processing during metastatic dispersal. Hence the morphology from thousands of cells need to be captured over an extended period of time. Furthermore, to account for the spatial effect of neighbouring cell, a potentially key factor in signal processing, cases of high cell density

also need to be considered. As a result, the large quantity of data coupled with the need to capture accurate and precise morphological details, for which all future analysis is based upon, means that the use of a CNN is the only option.

The chapter proceeds in the following manor. Firstly, the theory behind how a CNN segments an image is discussed as well as how it is trained. This theoretical understanding is then put into practice by deploying a network on series of phase contrast images. After the initial deployment the network is then retained and optimised. Finally, the chapter then concludes by discussing the optimisation results whilst also highlighting some features that are especially prominent in context of cell tracking.

2.2 Convolutional neural network

A convolutional neural network is a form of representation learning where the discriminatory features and characteristics within a dataset are automatically learned by the system (Bengio et al., 2013). These features are then used in tasks such as classification and segmentation (LeCun et al., 2015). The design of a CNN is such that a series of convolutional layers, often known as filters, pre-process an image and conduct the feature computation stage. This information is then passed to a neural network that performs the feature selection phase and thus the final classification (Khan et al., 2020). Note that throughout this chapter any reference to a neural network relates to a fully connected neural network unless otherwise stated.

2.2.1 Neural network

Graphically a neural network can be interpreted as a series of nodes that are grouped together to form a layer. Multiple layers are then stacked adjacent to one another with nodes in consecutive layers joined be a network of weighted edges (Figure 2.1) (Bengio, 2009; LeCun et al., 2015). The first and last layers are the input and output layers, with all intermediate layers referred to as hidden layers. The purpose of the hidden layer nodes is to transform the input data that they receive via a mathematical transform known as an activation function.

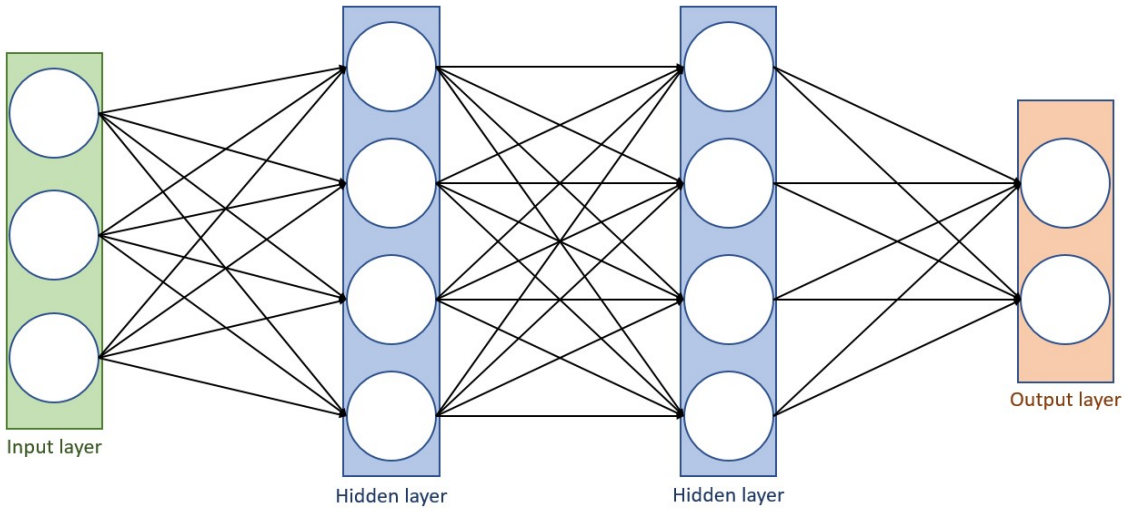


Figure 2.1: A graphical representation of a fully connected neural network.
The neural network has 3 nodes in the input layer (shown in green), 4 nodes in each of the 2 hidden layers (shown in blue), followed by 2 nodes in the output layer (shown in orange). The arrows indicate the weighted edges between nodes in adjacent layers.

The activation function is central in a neural network as it allows the network to detect non-linear patterns in the data (Webb et al., 2011).

$$Nodal\ output = f(\sum_{i=0}^{N} \overbrace{w_i * input_i}^{\substack{Weighted \\ input\ data}} + bias) \quad : \quad \begin{array}{l} f = activation\ function \\ N = number\ of\ input\ edges \end{array} \tag{2.1}$$

Neural network workflow

A neural network operates in a series of sequential stages. Firstly, a single numerical value is provided to each node in the input layer. For example, if the neural network is designed to classify cars by make the input data maybe engine size, weight, and top speed. This information is then passed through the connected edges of the network to the nodes in the second layer. The sum of the weighted input data, along with a bias, is then transformed by

the activation function of each node (Equation 2.1). The nodal output is then passed forward to the connected edges and the process is repeated across all of the hidden layers. A deep neural network can contain over 100 hidden layers meaning that this process is extremely computationally expensive (Bengio, 2009). Finally the nodes in the output layer represent the different classification classes i.e. the makes of car. The input value to each node in the output layer corresponds to the probability that the original data resides in the given class. The class with the highest probability is the final classification (Gardner and Dorling, 1998).

Activation function

The activation function is key to the network as it introduces non-linearity to the system which is essential when trying to understand complex mappings (Olgac and Karlik, 2011). In contrast to a mappings such as that between height and weight where a linear relationship is often sufficient. The mapping between data types such as the engine size and the make of a car are not necessarily as clear. The activation function is however helped by the presence of a bias as it allows the function to slide along the x axis and improve the model fit. In the context of linear regression the bias is akin to the intercept.

The non-linear nature of the activation function also has an important role in allowing the network to become deeper and thus leverage the multiple layers of abstraction. If the activation function is linear, irrespective of the number of layers in the network, then the system can be reduced to a linear mapping in a single layer. Hence the neural network would be akin to a linear regression model and as such it would lose its hierarchical power. In contrast, the same is not true with a non-linear function. Stacking multiple non-linear functions together allows the network to retain its hierarchical power and therefore detect features with a high discriminatory power. This framework can be further extended to allow each node to have its own activation function although in practice this level of complexity is often not beneficial.

2.2.2 Neural network training

A neural network learns the features within a data set through a process known as training. The goal of training is to minimise the error between the network prediction and the ground truth value, known as the loss function. The loss function is the error across all of the training examples for a given combination of parameters and hyper-parameters. Therefore, assuming that the training set is representative of the population, the location of the global minima on the loss function corresponds to the optimal combination of parameters and hyper-parameters in the network. If the loss function is visualised as a hilly landscape, then aim of training is to find the lowest point (Li et al., 2018). Hence the algorithm finds the lowest point on the function in the same way that a person finds the lowest point in a landscape, by walking downhill one step at a time.

Parameters

The edge weights and biases of a neural network are collectively referred to as the network parameters because they are iteratively adjusted after each epoch of training. An epoch is one complete pass through the entire training data set (Webb et al., 2011). However, due to the number of parameters and the non-linear relationship between layers the network's parameter space is often very large and complex. As a result, hundreds of epochs are required to adequately traverse along the loss function and find the global minima. The exploration itself is typically governed via a method of stochastic gradient descent (SGD) (Bottou and Lecun, 2004). SGD is an algorithm that steps along the loss function by taking a sample of points and calculating their lose. The gradient between each of the sample values is then calculated and the algorithm then steps in the direction of the largest negative gradient (Rumelhart et al., 1986), the steepest downwards slope.

Hyper-parameters

In contrast, the hyper-parameters such as the number of hidden layers or the activation function in a given layer are set at the start of each training schema (Aghdam and Heravi,

2017). As well as defining the architecture of the network the hyper-parameters also control how the SGD algorithm moves along the loss function. One of the most important hyper-parameters is the learning rate, the step size taken by the algorithm after each epoch (Murphy, 2012). If the learning rate is too small the algorithm will take a long time to converge and may also become trapped in a local minima. Likewise, if the learning rate is too large then the algorithm might miss the global minima altogether (Zeiler, 2012). Hence, to ensure that the learning rate, along with the other network hyper-parameters, is optimised correctly multiple models are trained and then compared to evaluate the performance of different model combinations (Goodfellow Ian et al., 2016).

2.2.3 Convolutional layers

In the previous example a vector of data was used to classify a vehicle. The elements of the vector corresponded to different features used to discriminate between different makes of car. However, in the case of an image, such features are not already known. A default option is to flatten the image into a single vector where each element corresponds to a given pixel. However, flatten the 2D matrix that represents an image means that input vector size can be considerable. A single 1080 x 1080 image results in a vector that contains over 1,000,000 elements (Aghdam and Heravi, 2017). Hence due to the number of weights in the network the model is prone to over-fitting.

Filters

CNNs combat this issue be leveraging the spatial information in an image to their advantage. CNNs utilise the knowledge that a given pixel is more strongly related to the pixels that are closer to it than pixels that are distant in the image (Krizhevsky et al., 2012). This enables a single image to be interpreted as a composition of multiple features i.e. an image of a car can be broken down into wheels, windows, and doors. This is achieved by convolving, or sliding, a filter across the image and extracting the different features. The filter itself is an $N \times N$ matrix where the elements within the filter are parameters that are estimated during training, similar to the weighted edges in a neural network. Once the filter has passed over

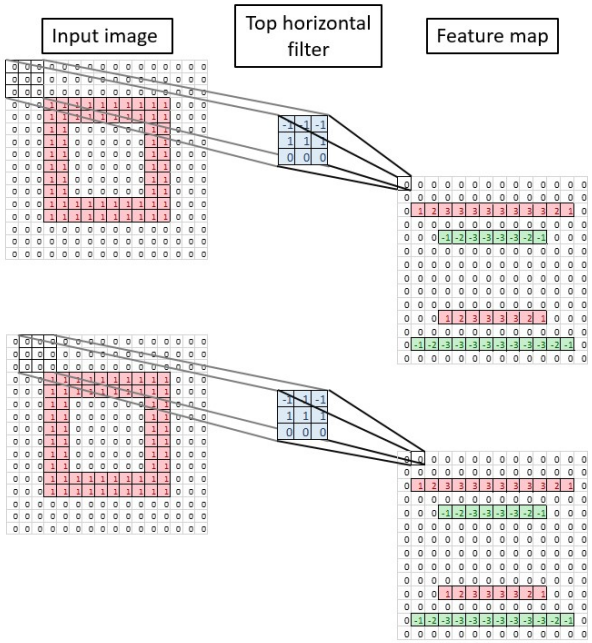


Figure 2.2: A top horizontal filter convolving across an input image.
The 16 x 16 input image contains a 10 x 10 square in the centre. The top horizontal 3 x 3 filters has a stride of 1 and maps to a single pixel in the 14 x 14 feature map. The red rows in the feature map indicate a positive result from the filter, a top horizontal row in the input image. The green rows in the feature map indicate a direct opposite result of the filter in the input image, a bottom horizontal row in the input image. The magnitude of the matrix elements in the red or green rows corresponds to fit of the filter to the region in the input image, the larger the magnitude the better the fit. The matrix elements of the filter are kept constant as it moves across the input image.

an image the subsequent output is referred to as a feature map. The feature map functions as a record containing the spatial location of a given feature within the original image (Figure 2.2) (Aghdam and Heravi, 2017). In the car example a filter might be used to extract cases where a wheel appears in an image. The feature map would therefore be the original image with only the wheels showing. Furthermore, by retaining the spatial structure of the image it means that individual pixels can be classified within the original image as well as the image as a whole, a process known as semantic segmentation.

In practice multiple filters are used in each convolutional layer to extract different features from the image (Figure 2.3) (Zeiler and Fergus, 2013). The initial layers contain low level filters that extract edges and curves. Whilst later layers contain filters that extract high level

features such as corners and shapes. The increase in filters does however mean that there is a corresponding increase in the number of feature maps that are produced at each convolutional layer (Figure 2.3) (Zeiler and Fergus, 2013). Once generated each of the feature maps are then transformed by a non-linear activation function before being passed to the next layer in the network. Similar to a neural network the activation function is critical as it allows the network to detect non-linear patterns within the feature maps. An example of a common activation function is the Rectified Linear Unit, ReLU (Glorot et al., 2011). The ReLU function operates by setting all negative values in the feature map to 0 whilst leaving all non-negative values untouched (Figure 2.3).

2.2.4 Leveraging spatial information

Sparsely connected

Retaining the spatial information in an image also has two other key benefits. Firstly, in contrast to a fully connected neural network, consecutive convolutional layers are only sparsely connected. Therefore, each pixel in a feature map is only connected to a small region of pixels within the input image (Sainath et al., 2013). This then reduces the number of parameters that need to be estimated during training and hence reduces the change of over-fitting.

Translational equivariance

Secondly, the parameter values within a given filter are estimated across the entire image. In a fully connected neural network a weighted edge is used once between a single input and output node. In a CNN the same mapping is used across nearly all of the input and output nodes, know as parameter sharing. The benefit of parameter sharing is that the network is equivariant to translation, a feature is detected regardless of its location within the image (Kondor and Trivedi, 2018). This is critical in a CNN because it means that the number of possible classes can be reduced dramatically. For example, rather than having 4 different

classes of a car for each quadrant of the image, the network only needs to retain a single class.

This is also an essential feature in the context of cell tracking as it means that the same cell is detected as it moves through space. Likewise, this also means in addition to classifying the image as a whole i.e is a cell present, the network can also record the spatial location of a cell within the image. This dual capability is achieved by using two independent neural networks where one handles the image classification and the other handles the object localisation. The object localisation network records the position of the object by using a boundary box that encompasses the perimeter of the object. Hence in the car example one network would classify the make of car whilst the other network records its position in the image e.g bottom left.

Pooling layers

Although the architecture of a CNN is dominated by convolutional layers the network is also punctuated with periodic pooling layers. The function of a pooling layer is to down sample the original input image into a smaller dimensional output. A common pooling function is the max output function that summarises a $K \times K$ area within the original image by its maximum value in that area (Krizhevsky et al., 2012) (Figure 2.3).

The benefit of a pooling layer is two fold. Firstly, the reduction in dimension means that there is a corresponding reduction in the number of parameters that need to be estimated within the network (Krizhevsky et al., 2012). Secondly, it makes the network invariant to translation (Graham, 2014). Translational invariance is similar to translational equivariance but with an important difference. The former means that a small deviation in the input value will map to the same output value. Whereas the latter means that a large deviation in the input results in the same deviation in the output. Thus in practice these two functions complement one another. The invariance criteria means even if there is a small change in the structure of an object at a local level it will still be classified correctly i.e. a car will be detected if there is a small dent in the body work. Whereas the equivariance condition allows the network to

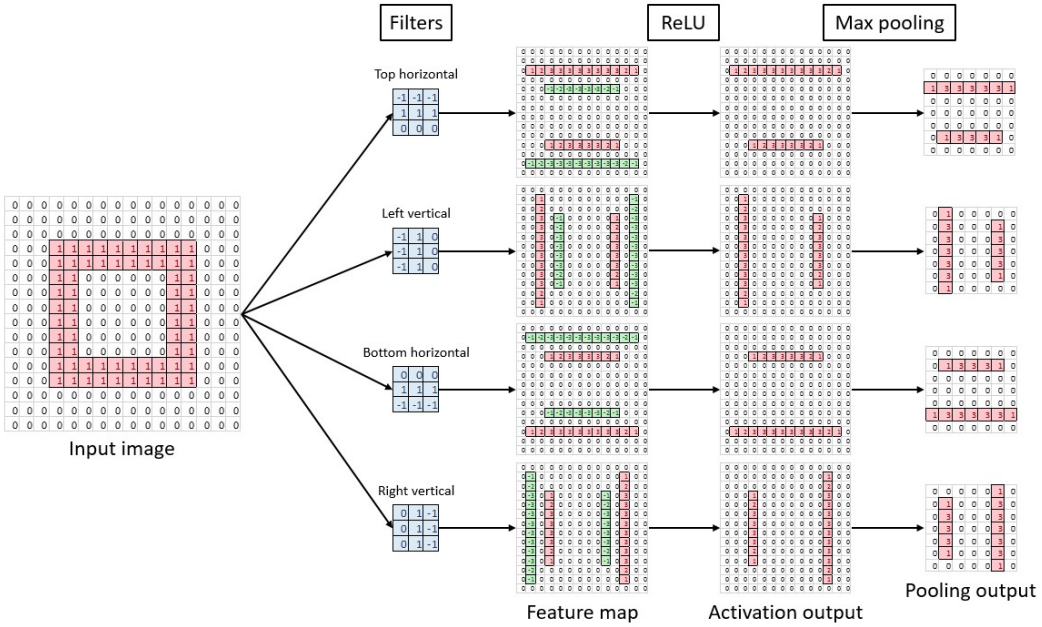


Figure 2.3: The workflow of a convolutional layer in a CNN.
A 16 x 16 input image with a 10 x 10 square functions as the input to the convolutional layer. The 4 filters with a stride of 1 are deployed to extract the 4 outer edges of the square, each of which has a length of 10, along with the 4 inner edges each with a length of 6. The feature maps are the same as Figure 2.2 where the red rows represent a positive match from the filter and the green rows represent the opposite filter match. The ReLU activation function then removes all negative matrix elements for the feature maps. Finally, the max pooling function down samples the activation output by taking the maximum value from a 2 x 2 region. The max pooling function has a stride of 2 and therefore reduces the 14 x 14 activation output to a 7 x 7 output.

detect the same features at different locations globally in the image i.e. a car is detected in the upper right or lower left corners of an image.

2.2.5 Region based CNN

The previous sections have outlined how a CNN classifies and localises a single object within an image. Yet in practice an image will often contain multiple instances of a given classes i.e. multiple cars within a given image. This variability presents a problem for a CNN as the length of the output layer is equal to the *number of instances* $\times$ *number of classes*. Hence if

the number of instances is not known in advance then the length of the output layer cannot be defined (Zhao et al., 2019).

Fortunately, region based CNNs overcome this issue by deconstructing the image into a series of smaller regions (Girshick et al., 2014). The smaller regions are then treated as individual input images before being reconstructed back into a full size image after segmentation. The virtue of this approach is that if the image is deconstructed into small enough regions then the number of instances within a region will always be either 0 or 1. In turn this allows the length of the output layer to be set equal to the number of possible classes within the image. Furthermore, through the addition of another neural network each instance within the image can also be segmented at a pixel level, known as instance segmentation (He et al., 2017).

Instance segmentation is critical in cell tracking as it enables phenomena to be investigated at a single cell level whilst in the presence of the broader cell population (Kimmel et al., 2018). Also, for logistical reasons a time-lapse video rarely records the entire migratory population. Instead, multiple geographical regions within the population are chosen as a sample and then their position remains fixed for the course of the video. This results in the number of cells that are present within the video changing through time. Furthermore, cells can divide or die during the course of a video which in turn alters the number of objects that are present at any one time.

2.3 Methods

All of the time-lapse data that was collected in this thesis was tracked using the semi-automated Usiigaci pipeline (Tsai et al., 2019). The pipeline combines a mask regional convolutional neural network, Mask R-CNN, (He et al., 2017) for segmentation with a particle tracker, Trackpy, (Allan et al., 2018; Crocker and Grier, 1996) for object linking (Figure 2.4).

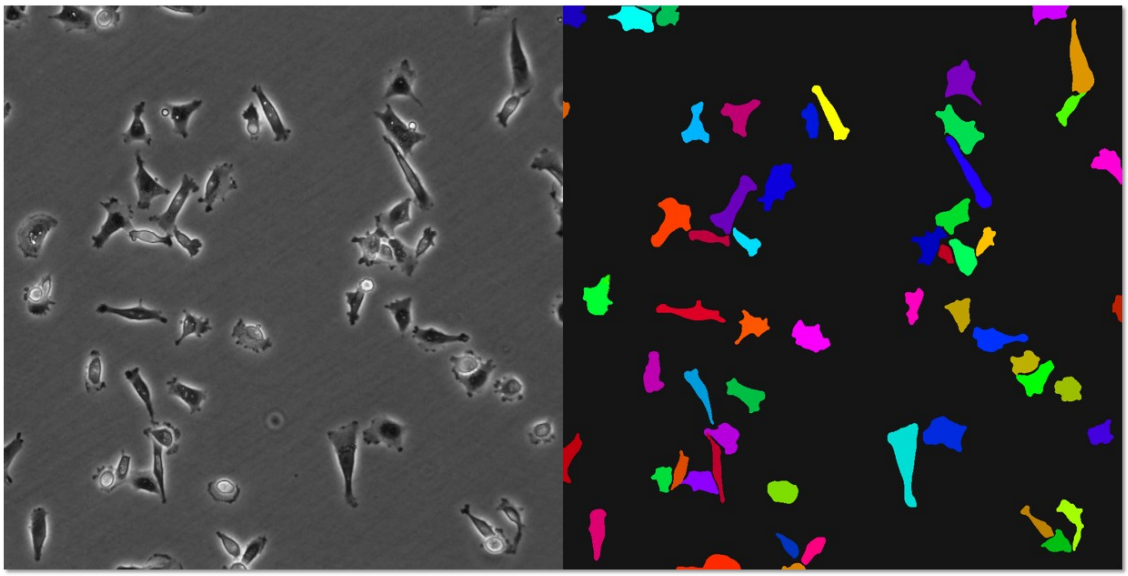


Figure 2.4: A phase contrast image segmented by Mask R-CNN.
The left hand image is a single 960 x 960 phase contrast image of MDA-MB-231 breast cancer cells. The right hand image is the corresponding 960 x 960 segmented output from Mask R-CNN that is used in future analysis.

2.3.1 Mask R-CNN

Mask R-CNN is widely considered to be the gold standard in instance segmentation as it combines high level precision with fast segmentation speed. The structure of Mask R-CNN can be broken down into 3 major stages (He et al., 2017):

1. ***Backbone**:*

   The backbone is a convolutional neural network that performs the feature extraction and is typically either ResNet50 or ResNet101, the suffix denotes the number of layers in the network. The ResNet architecture allows the networks to be extremely deep through the inclusion of "skip connections" that prevent the network from becoming saturated (He et al., 2016). In essence, skip connections allow the network to miss certain layers if the increase in depth is not providing an increase in performance. In theory this means that a deeper network will also perform at least well as its shallower counter parts. The final layer contains a feature map with dimensions 32x32x2048,

where 2048 corresponds to the number of individual feature map layers demonstrating the extreme degree of abstraction.

The backbone also includes a feature pyramid network that ensures that objects at multiple scales are detected. The feature pyramid network utilises the high level feature maps with strong semantic information, cases where large structures are detected, to reconstruct lower levels and ensure that smaller objects are also detected (Lin et al., 2017). This is critical in cell tracking as cell size can vary greatly and without the feature pyramid network smaller cells may not be detected. Furthermore, due to the sequential workflow of a CNN if a cell is not detected it cannot be segmented. As a result, the feature pyramid network is extremely important.

2. ***Regional Proposal Network****:*

   The region proposal network is a light weight neural network that selects different sized regions from the feature map and determines with a given probability whether the region contains an object (Ren et al., 2015). The classifier is binary and therefore if the probability exceeds a pre-defined threshold then the region is accepted. The regions themselves are quantified using boundary boxes that also record the region's position in the image. If the region does contain an object, a cell, then the region proposal network will make small adjustments to the boundary box location to improve its accuracy.

   The accuracy of the boundary box is quantified by the Intersection over Union, IoU (Rezatofighi et al., 2019). The IoU is the ratio of area covered by both the prediction and the ground truth compared with the total combined area covered by the prediction or the ground truth. An IoU of 1 indicates that the prediction has perfectly matched the ground truth. A high performance from the region proposal network is imperative. If the region proposal network does not suggest regions for all of the cells in the image, then a cell will be missed. As a result, the false negative rate is more pivotal than the false positive rate.

3. ***Headers****:*

The headers are 3 independent fully connected neural network branches that conduct the final: classification, boundary box localisation, and produce the pixel level segmentation mask for each instance. Importantly the loss function for the network is the combined loss for each of the 3 neural networks and therefore each branch needs to be trained effectively.

The classification branch operates in a similar way to the region proposal network classifier to determine the correct class for the object. However, in contrast to the region proposal network the classifier is discrete rather than binary. This allows different cell types to be classified within the same image, albeit all of the time-lapse data in this thesis uses a single cell type. Likewise, the boundary box localisation branch also follows a similar format to the region proposal network by adjusting the final boundary box to improve the localisation performance (Ren et al., 2015). Finally, the mask branch operates on a pixel by pixel manner to outline the shape of each cell. The thresholds for the boundary box refinement and the mask branch use the same IoU criteria mentioned above (He et al., 2017).

The Mask R-CNN network that is already in the Usiigaci pipeline has been pre-trained on the Common Objects in Context (COCO) data set before being further trained upon 50 phase contrast images. The COCO data set is an extremely large collection of manually annotated images that cover a broad spectrum of different scenarios. Due to its vast size, > 200,000 annotated images, the data set is often used as benchmark in computer vision competitions (Lin et al., 2015). The benefit of pre-training on a large general data set is to initialise the network parameters before re-training on a more specific, but often smaller, data set in which the network will be deployed. This two stage approach is extremely useful with deep CNNs as it helps to prevent the final network from over-fitting to the training data (He et al., 2017; Pan and Yang, 2010; Simonyan and Zisserman, 2015).

2.3.2 Performance evaluation

Once the final network has been trained its performance is then evaluated via the mean average precision, mAP (He et al., 2017). The mean average precision evaluates the networks precision at different recall levels. The precision measures how accurate the networks predictions are i.e. a precision of 1 indicates that all of the proposed cells are in fact cells whereas a precision of 0.5 indicates that only half of the predictions are correct. The recall measures how many of the cells are being detected within the image, a recall of 1 indicates that all of the cells in the image are being detected. For example, if the precision is 1 but the recall is 0.5 then only half of the cases are being detected but all of the detected cases are correct (Equation 2.2) (Everingham et al., 2015).

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{2.2}$$

In practice the precision is evaluated at multiple recall levels to determine the recall level at which the precision begins to deteriorate. Initially the precision will be high when the recall level is low, the object is only proposed if the network is nearly certain that the object is a cell. The precision then falls until nearly 0 when the recall approaches 1, everything is proposed as a cell and thus the network is not precise. The average precision is calculated as the area under the precision-recall plot and the mean average precision is the average precision across all classes (Goodfellow Ian et al., 2016). If only one class is present, as is the case in this thesis e.g. a cell, then the mean average precision is the average precision.

The threshold for a true positive is set prior the mAP being calculated and is based up on the IoU of the segmented mask. Historically the IoU threshold was set at 0.5, a prediction was accepted as a true positive if half of the segmented mask overlapped with the ground truth. However, the mAP is now more commonly evaluated over a range of IoU values to ensure that networks with a high level of precision are weighted more favourably (Huang

et al., 2016). In the context of cell morphology, it is critical to ensure that the network is segmenting cells with a high degree of precision as all future metrics are based upon these results.

2.3.3 Pre-trained performance

The performance of the current network within the Usiigaci pipeline was evaluated to determine whether re-training was necessary. To evaluate the network's performance 60 images were randomly chosen from two 72 hour time-lapse videos where an image had been taken every 2 minutes. The 4320 image data set was chosen as both time-lapse videos were recordings of MDA-MB-231 breast cancer cells, the same cell line that was used throughout this thesis. The length of both videos also allowed for multiple cell divisions to occur which ensured that some of the images were denser than others. The 60 images were then manually annotated and are referred to as the test set for the remainder of this chapter. The manual annotations involved drawing around the outline of every cell in each of the 60 images to produce a ground truth mask. The average mAP score was then calculated across 10 IoU thresholds ranging from 0.5 - 0.95 (Huang et al., 2016).

The current Usiigaci trained network scored an average mAP across 2 repeats of 0.0389. This is extremely low considering the pre-trained network from the COCO data set scored a mAP of 0.3110. Nevertheless, the poor performance is not necessarily a surprise. Firstly, the Usiigaci training data is from a different cell type to the test set. Secondly, there are large differences in the image properties between the training and test data as well as a difference in the number of cells. In all, the poor mAP result confirms that the network needs to be retrained prior to deployment.

2.4 Results

2.4.1 Retrained performance

Training setup

The network was initially re-trained using the same protocol as the Usiigaci network, referred to as the baseline for the remainder of this chapter. The protocol used 50 manually annotated images to train the network with a 90:10 split for the training and validation subsets respectively. In contrast to the independent test set that is only seen by the network once for the final network evaluation. The training data is seen by the network throughout training and has two subsets: training and validation. The training subset is used to estimate the network parameters during each epoch. Then at the end of each epoch the validation set is used to independently evaluate how well the network has performed. Keeping the validation set independent from the training set is vital as it helps to prevent the network from over-fitting to the training data (Goodfellow Ian et al., 2016). To ensure that the terminology is clear for the remainder of this chapter the training data refers to the training set and the validation set combined.

The re-trained network kept all of the hyper-parameters the same as the baseline as well as the training schema. The training schema consisted of 100 epochs of training on the network headers followed by 100 epochs on the full network. Across the 200 epochs the learning rate remained fixed at 0.001. Training the network in parts is sometimes used in cases where the training data is small. Optimising the headers typically requires less data than the network backbone and therefore training in sections can help to prevent over-training (Zhang et al., 2020).

After having repeated the same protocol as the baseline the new training data yielded a dramatic improvement in the network's performance, the mAP score increased to 0.541. The marked increase in performance highlights the sensitivity of the network to the data that it is trained upon. Hence ensuring that the training data accurately represents the population in which the network is going to be deployed in is paramount.

Practical performance maxima

Although the increase in performance is sizeable, it does not necessarily mean that an optimal model has been found. A perfect segmentation has a mAP score of 1. Yet in practice no CNN will achieve this as the network is limited by the Bayes error, the unaccountable stochastic error within the system (Tumer and Ghosh, 1996). Likewise a sub optimal performance will not necessarily have any impact on the subsequent biological interpretations.

To generate a practical performance maxima 10 images were randomly chosen from the 60 image test set and re-annotated. The 10 re-annotated images were then evaluated against the same 10 images from the original test set, as if they had been segmented by a trained network. The re-annotated images recorded a mAP score of 0.658 and therefore this will be used as a practical maxima for the network to reach.

Training schemas

To determine whether the network can reach the practical performance maxima different training protocols were executed and then evaluated on the test set. The different protocols involved independently varying the training data size, validation split, and two of the hyper-parameters.

The 14 different network configurations were also each tested with 4 different training schemas (Figure 2.5). The different training schemas were deployed to evaluate the benefit of training the headers disjoint from the rest of the network, S1 vs S2, and to compare the use of a fixed vs variable learning rate that decreases over training, S1/S2 vs S3/S4 (Figure 2.5). The motivation for a variable learning rate is to allow the network to explore the landscape globally at the start of training, before then narrowing on a particular region as the training progresses (Bengio, 2012). In total 56 different networks were trained and by the end of each training period the loss function had converged. This process was then repeated and the average of the two repeats was recorded as the final mAP for each network combination for each of the 60 test images.

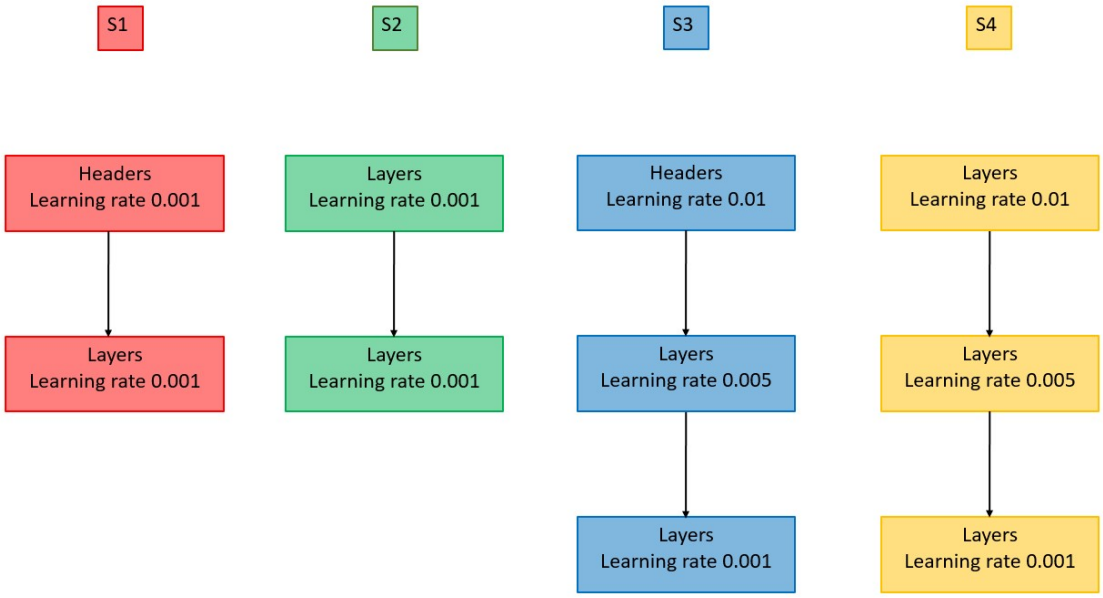


Figure 2.5: An overview of the 4 different training schemas used for optimisation. S1 has a fixed learning rate of 0.001 with 100 epochs initially on the headers followed by 100 on all layers. S2 has a fixed learning rate of 0.001 with 100 epochs on layers followed by another 100 epochs on all layers. S3 has a variable learning rate starting at 0.01 on the headers for 100 epochs, followed by 100 epochs on all layers at a learning rate of 0.005, and finally another 100 epochs across all layers at a learning rate of 0.001. S4 has the same decrease in learning rate over 300 epochs as S3 but it runs across all layers of the network.

2.4.2 Training data structure

Training data size

The first comparison tested whether an increase in training data caused a corresponding increase in model performance. The training data size was increased from 40 to 240 images with a 40 image increase at each interval. The validation split remained constant with 90% in the training set and 10% in the validation set, the same as the baseline. All hyper-parameters were also kept the same as the baseline and each training data size was deployed across all 4 training schemas.

A linear mixed model was then used to evaluate whether the relationship was significant. The model was such that the mAP score was dependent on the training data size and the difference between images in the test set was controlled with a random effect. Finally, the

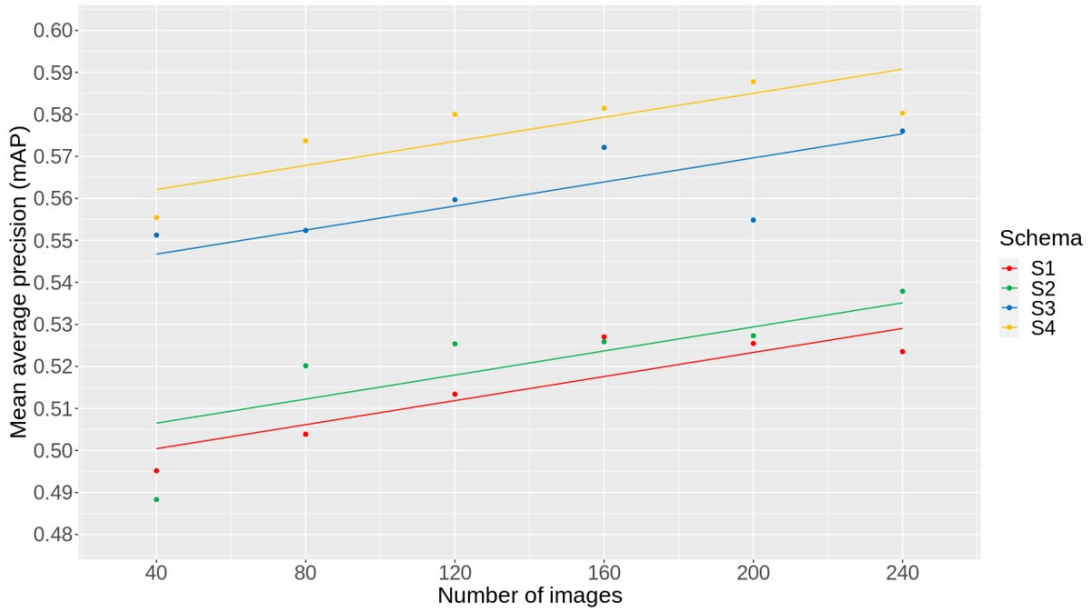


Figure 2.6: A plot of the mAP score against the number of images in the training data. The training data size was increased from 40 to 240 images with a 40 image increase at each interval. The data points represent the average mAP score across the 60 image test set for each of the different training data sizes. The straight parallel lines correspond to the significant parameter values of the linear mixed model that was fitted to the data. The intercepts were allowed to vary between each training schema and were all found to be significantly different from one another at a 5% level. The slope parameter ($\beta = 1.43 \times 10^{-4}$) was also found to be significantly different from 0 at a 5% level.

model intercepts were allowed to vary across the 4 training schemas but the slope remained fixed.

All 4 intercepts were found to be significantly different from 0 and from one another at a 5% level (N= 1440). The fixed learning rate schemas, S1 and S2, had intercept values of 0.495 and 0.501 respectively. Whilst the variable learning rate schemas S3 and S4 had significantly larger intercept values of 0.541 and 0.556 respectively (Figure 2.6).

The slope parameter ($\beta = 1.43 \times 10^{-4}$) was also found to be significant at the 5% level ($p < 2 \times 10^{-16}$, N = 1440) (Figure 2.6). Thus, meaning that a 100 image increase in the training data size causes a 0.014 increase in the mAP score. The significance also means that the largest possible training data size should be used to maximise the mAP score. Henceforth, all remaining network evaluations used a training data size of 240 images. Finally, the fixed

effects of the mixed model explained a significant proportion of the variation in the mAP scores with a marginal $R^2 = 0.145$.

Validation split

Next, the validation split was increased to evaluate whether training data size and structure had a significant effect on the model performance. The training data size was fixed at 240 images but the validation split increased from 10% to 25% in 5% intervals. All hyper-parameters were kept the same as the baseline and each validation split was deployed across the 4 different schemas.

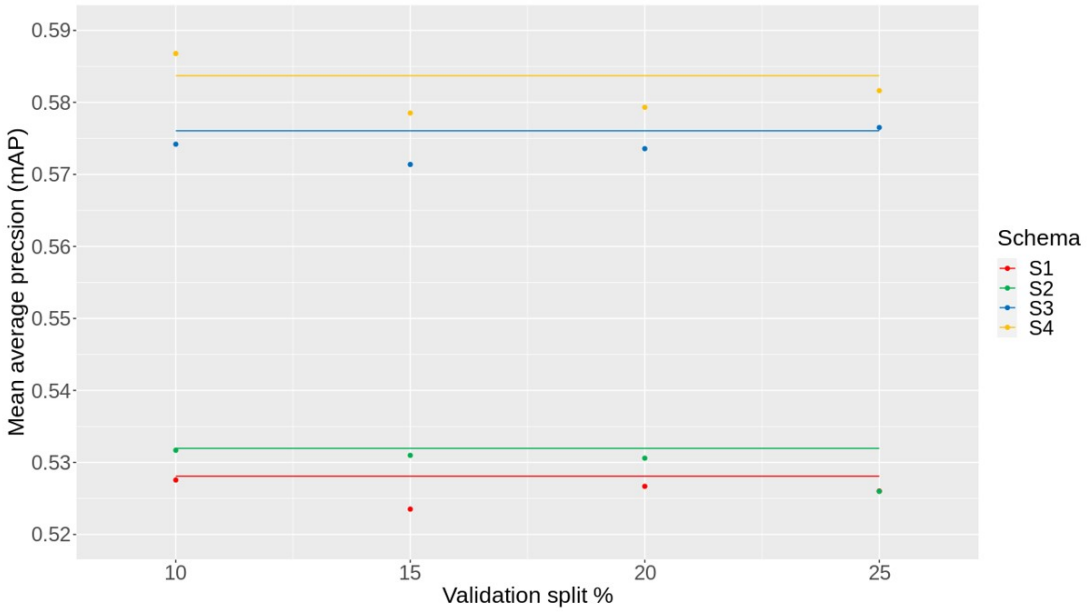


Figure 2.7: A plot of the mAP score against the validation split in the training data. The validation split was increased from 10% to 25% in 5% intervals. The data points represent the average mAP score across the 60 image test set for each of the different validation splits. The straight parallel lines correspond to the significant parameter values of the linear mixed model that was fitted to the data. The intercepts were allowed to vary between each training schema and were all found to be significantly different from one another at a 5% level. The slope parameter was not significantly different from 0 at a 5% level and thus the slope value was set to 0.

A linear mixed model was also used to evaluate the significance of increasing the validation split on the network performance. Similar to the training data size model the mAP

score was dependent on the validation split and the difference between images in the test set was controlled with a random effect. Likewise, the intercepts were allowed to vary across the 4 training schemas but the slope remained constant.

All 4 intercepts were found to be significantly different from 0 as well as from one another at the 5% level (N = 960). The variable learning rate schemas, S3 and S4, had intercept values of 0.576 and 0.584 respectively. These both outperformed the fixed rate learning schemas which had mAP scores of 0.528 and 0.532 for schemas S1 and S2 respectively (Figure 2.7).

However, in contrast to the training data size, the slope parameter was not significant in the validation split. There was no significant correlation between an increase in the validation split and an increase in the mAP score (Figure 2.7). As a result the validation split remained at 10% for the remaining network comparisons.

2.4.3 Hyper-parameter optimisation

The two previous sections evaluated the effect of changing the training data size and structure on the network's performance. However, another kept aspect that is under the control of an operator is the choice of hyper-parameters that are used during training. The hyper-parameters control the networks structure as well as dictating how the algorithm moves through the parameter space during training. The total number of possible hyper-parameters are vast and as such not all combinations could be evaluated. Nevertheless, two highly influential hyper-parameters, the backbone and gradient clip norm, were tested.

ResNet backbones

The backbone, as detailed in Section 2.3.1, controls the number of convolutional and pooling layers that are used in the model as well as the structure of the layers. The two backbones that have been evaluated in this section are ResNet50 and ResNet101. The suffice details the number of layers, 50 and 101, whilst the ResNet prefix indicates the presence of "skip connections". Skip connects allow layers to be missed during the back-propagation phase of training, the process by which parameters are optimised (He et al., 2016). This in turn helps to mitigate the issue of vanishing gradients that plague deep neural networks. A vanishing

gradient refers to when a weight gradient becomes extremely small during training. An extremely small gradient means that the network training becomes extremely slow, or worse still training may halt altogether as the network does not know which direction to move in (Goodfellow Ian et al., 2016). The inclusion of skip connections however allows the network to miss a layer during training if the gradients become too small. In practical terms this means that adding additional layers to a network should result in the same or better performance compared to shallower networks.

Gradient clip norm

Gradient clipping also helps with gradient related issues during training. However, in contrast to the skip connections, gradient clipping deals with the opposite issue when gradients become extremely large. During training, the SGD algorithm can end up moving through areas of the landscape that are rough and jagged. In turn this means that steep descents and cliff like edges are often present. When combined with a large learning rate this can produce large gradient changes causing dramatic weight adjustment within the network. This then means that the algorithm can end up overshooting the minima or the network may become numerically unstable causing the training to terminate (Bengio, 2012). Keeping the learning rate low reduces the prevalence of gradient explosions but they can still occur. Gradient clipping however helps to prevent this by rescaling the gradient. Rescaling the gradient has a similar effect to reducing the learning rate of the algorithm by reducing the step size and preventing the minima from being missed (Kim et al., 2016). In this section a gradient clip norm of 10 was compared against a clip norm of 5 in the baseline. The training data size was set at 240 images and the validation split remained at 10%.

Hyper-parameter results

A Kruskal Wallis test found that ResNet101 backbone significantly outperformed the ResNet50 backbone irrespective of the gradient clip value or the training schema ($p < 2.2 \times 10^{-16}$, N = 960) (Figure 2.8). Although the difference is striking the result is not unexpected. It confirms the information in the previous paragraph that a deeper network

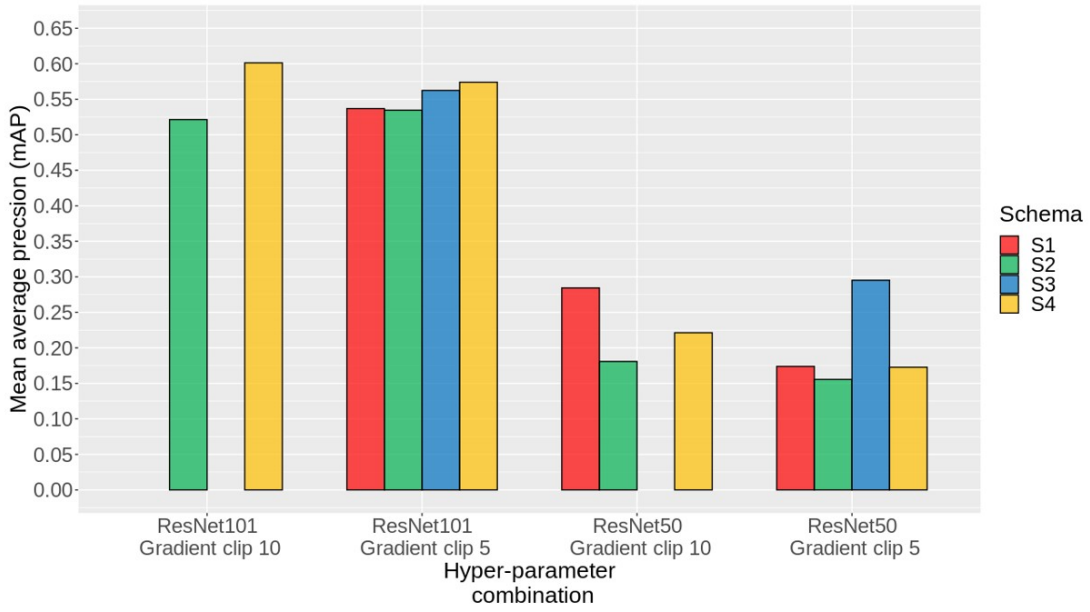


Figure 2.8: A plot of the mAP score against 4 different hyper-parameter combinations. The ResNet backbones were compared at two different level as well as the gradient clip value was set at either 5 or 10. The height of the bar chart represents the mAP score of each hyper-parameter combination for each training schema. The ResNet101 backbone performed significantly better than ResNet50 across both gradient clip values regardless of the training schema. Training schema S4 performed better than training schema S2 within the ResNet101 backbone at both gradient clip values. Finally, the highest mAP score was achieved with the ResNet101 backbone, a gradient clip of 10 and the S4 training schema.

will always perform at least as well as shallower network. However, it was still important to verify the result on a relatively small training data size of 240 images. Due to the extreme performance difference the remaining gradient clip comparisons focus solely on the ResNet101 backbone.

The ResNet101 backbone had 8 different network combinations to test. However, two of the training schemas with a gradient clip of 10, S1 and S3, become numerically unstable and did not finish. This maybe due to an interaction between the larger gradient clip norm and the 100 epochs of training that solely focuses on the headers in both schemas S1 and S3. The same behaviour was also seen in schema S3 when the ResNet50 backbone was combined with a gradient clip of 10. As result, the two different gradient clip values were compared with the ResNet101 backbone and just for schemas S2 and S4.

A mixed model analysis of variance (ANOVA) was used to evaluate the two gradient clip norm values. The gradient clip value and training schema were set as fixed effects along with the interaction. The random effect controlled for the difference between images in the test set. The gradient clip norm and training schema were both significantly different from 0 ($p = 9.337 \times 10^{-3}$ and $p < 2.2 \times 10^{-16}$ respectively, N = 240) as well as the interaction ($p = 1.164 \times 10^{-12}$, N = 240). The fixed effects of the model also explained a significant proportion of the variation in the mAP score with a marginal $R^2 = 0.187$.

The significance of the interaction term implied that the effect strength of the gradient clip norm varied between training schemas. To test which of the 4 network combinations were different from one another a pairwise Bonferroni multiple comparison test was used. In turn all 6 of the pairwise comparisons were found to be significantly different from one another at a 5% level. The highest mAP score with a value of 0.601 was produced with a ResNet101 backbone, a gradient clip norm value of 10, and a variable learning rate schema that trained across all layers. This was an 11.1% mAP increase relative to the baseline performance and it meant that 91.3% of the practical performance maxima had been achieved. This model configuration is referred to as the alpha configuration for the rest of this chapter.

2.5 Discussion

Automated segmentation is a major obstacle in computational cell tracking (Caicedo et al., 2017). The non-rigid nature of a cancer cell means that it is extremely difficult to define as a target and the low contrast between the cell and the background further complicates the issue. However, segmentation performance has been increased dramatically with the introduction of convolutional neural networks (CNNs) (Asgari Taghanaki et al., 2020). CNNs have been able to capture an unprecedented level of precision which has in turn allowed more complex features to be investigated. Furthermore, they offer a higher throughput capability due to their increased level of autonomy and lack of sustained operator input.

Nevertheless, full transferability has yet to be achieved. Thus in the majority of cases a network will need to be re-trained on a specific sample of data to account for the variability

between research groups. Thus, as a minimum, a high level understanding of the basic factors that influence a CNNs performance is necessary to ensure that the best possible results are achieved.

2.5.1 Training data importance

One of the most important factors in a CNNs performance is the data that the network is trained upon. The training data is used to estimate the network's parameters which in turn control how well a network will perform during segmentation. Time and attention should therefore be given to the quality of the training data that is used and to ensure that it is representative of the target population. Also, specific care should be given to any manual annotations that are performed as an erroneous error can have a substantial negative impact on the network's performance (Moen et al., 2019). Once the training data has been accurately collected and curated the next issue is to determine how much data is needed.

Optimal training data size

The error of a CNN, as with any model, is expected to decrease as the sample size increases. This explains why an increase in training data caused a significant increase in the mAP score of the network (Figure 2.6). The increase in training data means that a larger proportion of the population is now known. The larger proportion of known information means that the network error decreases which in turn causes an increase in the mAP score, 1 - error = mAP score. Yet, despite the significant relationship it is still not clear how much training data is needed. Furthermore, it would be incorrect to assume that the linear relationship will continue indefinitely and that increasing the training data size will always be beneficial.

Theoretically the network error will decrease until it is asymptotic with the Bayes error. Hence, the model performance will increase until it reaches an inflection point and begins to flatten. The location of the inflection point is important because it defines how much training data is needed to achieve an optimal network performance (Goodfellow Ian et al., 2016). Although the inflection point was not found in this analysis the model fit the data best when the exponent was equal to 1 rather than with a reciprocal term. This therefore suggests that

the relationship was not beginning to flatten. Thus, implying that considerably more than 5% of the image population needs to be sampled to reach an optimal network performance. As a result, the total training data size will most likely not be an active decision from the operator but rather it will be constrained by other external factors. The best practice approach should therefore be to collect as much training data as possible with the knowledge that any tractable increase in training data size will be beneficial to the model performance.

2.5.2 Training algorithm

An alternative approach is to improve the training algorithm through optimising the hyper-parameters in the model. One of the most important, but also most difficult, hyper-parameters to optimise is the learning rate. The learning rate needs to be large enough to allow for adequate exploration of the parameter space, but yet small enough to converge onto a minima when it has been found. Striking a balance between these two competing objects is an active area of research with a variety of different approaches. Yet broadly all of the approaches centre around the idea that the learning rate should vary over the training duration.

Adaptive learning rate

The benefit of a variable learning rate is evident across all of the training comparisons where schema S3 and S4 consistently outperform schemas S1 and S2, albeit with the exception of the ResNet50 backbone. Likewise, the alpha configuration also used the S4 variable learning rate schema (Figure 2.5). A further extension upon schema S4 is to allow the training process an even higher degree of autonomy such that it evaluates its own performance during training and then changes the learning rate accordingly. This is referred to as an adaptive learning rate as the learning rate changes dynamically during training rather than being fixed at the start of training (Kingma and Ba, 2014). An example of this approach can be seen with the AdaGrad algorithm that scales the individual parameter updates proportionate to the sum of the inverse historic gradient values (Duchi et al., 2011). Practically this means that training favours a direction that has a gradual slope rather than solely picking the steepest gradient. The motivation being that the algorithm will find a reliable path that descends towards a global

minima and stick to it rather than bouncing between ridge lines. Adaptive learning rates are often a popular chose when training a CNN because they tend to converge faster than stochastic gradient descent (Zeiler, 2012) and often more forgiving to optimise. In turn this is especially useful when working with a large training set size that might be prohibitively expensive to evaluate for each epoch of training.

2.5.3 Generalisation performance

Although appealing at first glance adaptive learning rates should also be approached with caution. There is evidence to suggest that the use of an adaptive learning rate might mean that the final solution is not as generally applicable as that gained via stochastic gradient descent (Keskar and Socher, 2017). This might not be an issue if the training data is extremely large or highly representative but in context of cell tracking it has important implications.

Time-lapse videos are commonly recorded in a sequential manner, often with large gaps between each recording. This is a product of the experimental workflow by which cells are cultured and is often unavoidable. Likewise results from previous experiments are often used to inform the decisions of future experiments and thus a sequential workflow can be beneficial. The downside of this is that the total population size grows through time rather than remaining fixed. Thus, initially the training data set is a random 5% sample of a single experiment. However, as more experiments are performed the sample becomes less representative of the population. Therefore, a high degree of generalisation is essential to ensure that the network is only re-trained for a given lab rather than needing to be re-trained for each experiment. As a result, stochastic gradient descent was kept as the only training algorithm through all training combinations.

Finally, to maximise the general applicability of the alpha configuration, multiple models where trained independently and then they were combined for the final evaluation. This is known as an ensemble approach. The motivation for an ensemble approach is that each model will make slightly different mistakes, model $M1$ will detect slightly different features to model $M2$ (Chandra and Yao, 2006). Hence when combined together the average model performance will benefit from the independent capabilities of each model and thus the mAP

score will increase. To determine whether an ensemble approach is beneficial a single alpha configured model was evaluated against the combined output of the single model plus another 3 independent alpha configured models. Although the training configuration remained constant across the 4 models the training images were randomly divided into the training and validation sets. This means that each of the models did not have exactly the same images in their validation set, and therefore different models were able to focus on slightly different features. After evaluating the two approaches the ensemble approach was found to have a 2% higher mAP score compared to the single model. This confirmed the use of a model ensemble and therefore all subsequent time-lapse data in this thesis were segmented by 4 independently trained alpha configured CNNs.

To summarise, cell tracking is a critical pre-processing stage that is essential when analysing time-lapse videos. In this chapter a variety of different automated cell tracking approaches were discussed before a convolutional neural network was chosen. The convolutional neural network was then optimised with a specific focus on the training data quantity and structure. In turn the trained convolutional neural network was then used for all subsequent cell tracking throughout the remainder of this thesis.

Chapter 3

A phenotypic switch in the dispersal strategy of cells selected for colonisation

3.1 Introduction

Metastasis is a form of long-range dispersal (Amend et al., 2016; Tissot et al., 2019) and central to understanding how cancers metastasise is understanding how cells migrate (Paul et al., 2016; Wells et al., 2013). During migration, as cancer cells become more invasive, they adopt an altered morphology, typically taking on elongated shapes characteristic of epithelial-mesenchymal transition (EMT) (Cowden Dahl et al., 2009; Odenwald et al., 2013). This change in cellular morphology is an important marker of migratory state (Prasad and Alizadeh, 2018; Wu et al., 2020). As result, quantitative measures of cell morphology taken from static images have been shown to effectively differentiate between cancer cell lines with high and low metastatic potential (Alizadeh et al., 2016; Lyons et al., 2016). Yet, there are important aspects of migratory behaviour linked to metastasis that cannot be captured solely from static images. This is especially evident when investigating dynamical behaviours such as signal processing ability.

Challenges of metastasis

To successfully metastasise a cell is required to navigate through a series of sequential steps known as the metastatic cascade. The cascade begins with a cell escaping from the primary tumour before then migrating through the extracellular matrix towards a nearby blood vessel. Once at a blood vessel the cell must then intravasate into the blood before it is carried around the body to a distant site. After reaching a distant site the cell then needs to extravasate from the blood and invade into the foreign tissue. Finally, the cell must re-initiate aggressive proliferation and colonise the distant site by forming a secondary macroscopic tumour (Valastyan and Weinberg, 2011).

In addition to the cellular changes needed for metastatic success there are also a host of environmental changes needed for a cell to metastasise (Shieh, 2011). This is evident from the onset of cellular dispersal where nearby collagen fibres are straightened perpendicular to the tumour boundary (Provenzano et al., 2008). The straightened fibres then act as a pathway for future migrants in turn improving their migratory success (Wershof et al., 2019). This dynamic cell-environment interplay continues throughout the metastatic cascade (Yuan, 2016) and highlights the importance of the surrounding micro-environment in shaping metastatic progression.

Morphological behaviour distant site colonisation

Successful colonisation of a distant site, the final and rate-limiting step of metastasis (Massagué and Obenauf, 2016), requires navigation through the unpredictable tumour microenvironment (Clark and Vignjevic, 2015) as well as the novel environment at a the distant metastatic site (Valastyan and Weinberg, 2011). In both stages success is achieved, in part, by the cell's capacity to detect and respond to changes in the environment (Costa-Silva et al., 2015; Peinado et al., 2011; Psaila, Kaplan, Port, and Lyden, Psaila et al.; Sceneay et al., 2013). Therefore, cells capable of distant site colonisation would be expected to have an altered signal processing ability and be more reactive to environmental change. As a result, an altered degree of morphological change might be expected in colonising cells. Furthermore, morphological change is expected to be positively correlated with migration

speed in successfully metastasising cells, because a faster-moving cell will experience a greater degree of environmental variation over a given time period, and therefore change its morphology more rapidly in response.

To test this hypothesis the chapter proceeds by first detailing the experimental selective pressures that were used to evolve three populations of cells each of which correspond to a separate stage of the metastatic cascade (Chaffer and Weinberg, 2011; Pantel and Brakenhoff, 2004): escape from the primary tumour, invasion of foreign tissue, and distant site colonisation. The chapter then explains how Zernike moments can be used to quantify the morphology of each individual cell within the evolved populations. This morphological information is then modelled in response to the speed of migration and the distance to neighbouring cells. Finally, the chapter concludes by discussing the morphological dynamics of each population in light of the selective pressure that were applied.

3.2 Data collection

To evaluate the precise phenotypic changes that are associated with metastatic success, it is preferable to compare replicate populations of cells that differ only in their ability to metastasise. Clinically this is often not possible as a variety of different selective pressures tend to exist within a tumour. However experimental evolution, a technique more commonly seen in microbiology, can be used to generate such cancer cell populations (Sprouffske et al., 2012; Taylor et al., 2013).

3.2.1 Evolved population summary

Starting with a population of MDA-MB-231 breast cancer cells three separate selective regimes were applied each of which was designed to be similar to the ecological pressures experienced whilst traversing the metastatic cascade (Valastyan and Weinberg, 2011). Two biological replicate ancestor populations were also frozen at the start of the experiment to act as a control for comparisons with the evolved lines (Figure 3.1) (Kawecki et al., 2012).

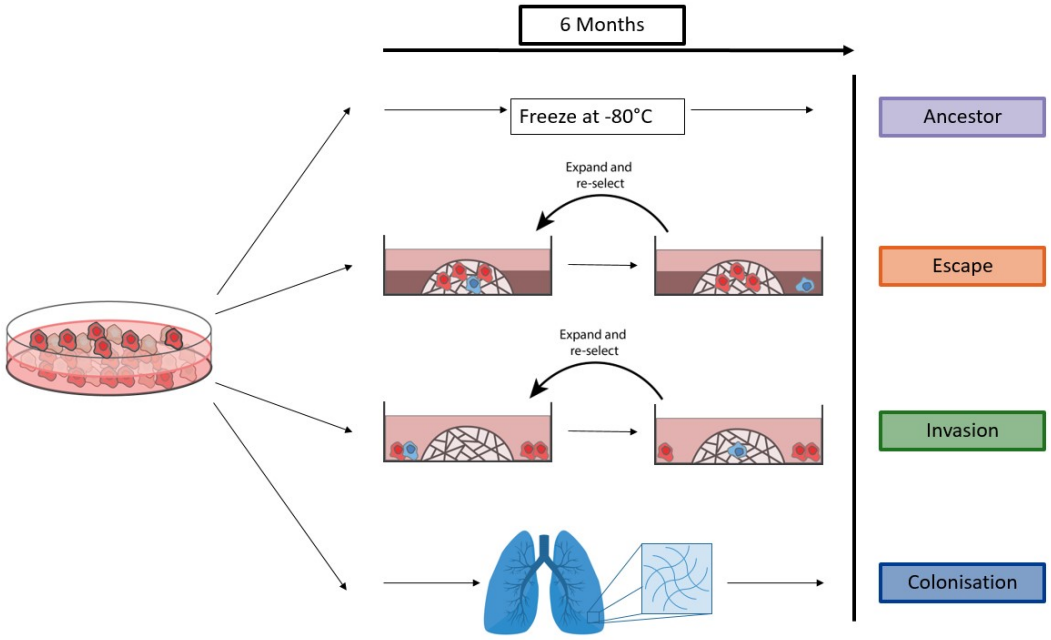


Figure 3.1: Experimental evolution of cancer cell populations.
Ancestor populations were kept frozen throughout. Escape populations were placed in a high density collagen matrix surrounded by a low density outer collagen ring: after 10-14 days cells that had escaped into the outer ring (shown in blue) were released, expanded and reseeded back into a new high density collagen core; this process was repeated 7 times over the course of 6 months. Invasion populations were seeded around a Matrigel island; after 7 days cells that had invaded the Matrigel (shown in blue) were released, expanded and reseeded around a new Matrigel island this was repeated 15 times over the course of 6 months. Colonisation populations were seeded onto a piece of decellularized rat lung which acted as a novel scaffold for colonisation and left to establish for 6 months. Four replicate lines were maintained for each treatment.

Escape populations

The escape populations (Figure 3.1) were selected by tightly packing cells into a high density core of collagen and then allowing them to escape outwards into a low density collagen outer ring (Keeton et al., 2018). After 10-14 days the cells that had escaped into the outer collagen ring were recovered from the matrix, expanded, and then seeded back into a new collagen escape assay, completing one round of selection. In total, 7 rounds of selection were applied to each of the four biological replicate escape populations. The high density collagen core and the low density outer collagen ring were both three-dimensional (3D) culture environments designed to be similar to those experienced during tumour dissemination.

Invasion populations

The invasion populations (Figure 3.1) were selected following a similar protocol to the escape populations whereby repeated consecutive rounds of selection were applied. In contrast to the escape assay, however, cells moved from a 2D to 3D environment, similar to the change in environment experienced during the arrest of a cell at a distant site. The cells were seeded around the outside of a Matrigel island - a synthetic basement membrane matrix widely used in cell culture - and left to invade. After 7 days the cells were collected from the Matrigel, expanded and seeded around the outside of another Matrigel island. This process was repeated 15 times for each of the four biological replicate populations over the course of the 6 month experiment.

Colonisation populations

The colonisation populations (Figure 3.1) were selected by culturing cells on a piece of decellularized rat lung, which acted as a scaffold for growth similar to that experienced by cells colonizing a distant site (Keeton et al., 2018). The protocol involved cells being seeded onto a decellularized scaffold and left to colonize over a 6 month period. Decellularized tissue is generated by removing all cells from a piece of tissue such that only the extracellular matrix is left. At the end of the experiment cells were released from the scaffold, ensuring that the population represented cells from within the tissue core as well as the edges. Again, this selection was applied to four biological replicate populations.

Finally, all twelve experimentally evolved cell populations were frozen and then thawed alongside the ancestor populations prior to experimental analysis. This step ensured that any selective pressure from the freezing-thawing process was constant across all treatments and replicate populations.

3.2.2 Time-lapse microscopy

Once thawed the cells were then placed onto 2D tissue culture plates and their migration was recorded over a 12-hour period, with images taken at two-minute intervals. This resulted

in 11,880 phase contrast images being collected over a total of 33 time-lapse videos. The 2D plastic environment was intentionally chosen as a neutral testing environment and to ensure that the morphology could be clearly seen without the use of fluorescent tags, a factor that might have applied an additional selective pressure (Liu et al., 1999). Further technical details related to the experimental assays and exact microscope settings can be found in Appendix A.

After the time-lapse videos had been recorded, the cells were then tracked through the semi-automated Usiigaci pipeline (Figure 3.5A) (Tsai et al., 2019) which had been trained with an alpha configuration as detailed in Section 2.4.3. The segmented morphology of each cell was then manually checked in every frame to detect any possible errors i.e a cell had divided, been mis-identified or incorrectly segmented. Furthermore, 30 minutes before and after a cell division were excluded to remove any rounded morphologies typical of a cell dividing (Cooper and Hausman, 2000; Théry and Bornens, 2006). Finally a cell needed to appear in at least 30 frames and be present for at least 75% of the trajectory to be included in any further analysis. A total 813 cells were tracked across the 33 time-lapse videos.

3.3 Quantifying morphology

Once tracked, the morphology of each cell needed to be quantified. Whilst a wide variety of different morphological measure exist, they can be broadly grouped into two main categories; descriptive measures and basis function expansions (Prasad and Alizadeh, 2018). Descriptive measures include metrics such as the cell perimeter, area, and aspect ratio. They are classified as descriptive because they measure a given feature that is then used to summarise the entire morphology of a cell (Pincus and Theriot, 2007). As a result, descriptive measures tend to have a straightforward biological interpretation and they are often easy to record. However, because they only measure a single feature are also limited by the prior belief of the operator. Hence not all of morphological variation is necessarily captured. Furthermore, descriptive measures tend to be highly correlated with one another which can subsequently restrict further analysis.

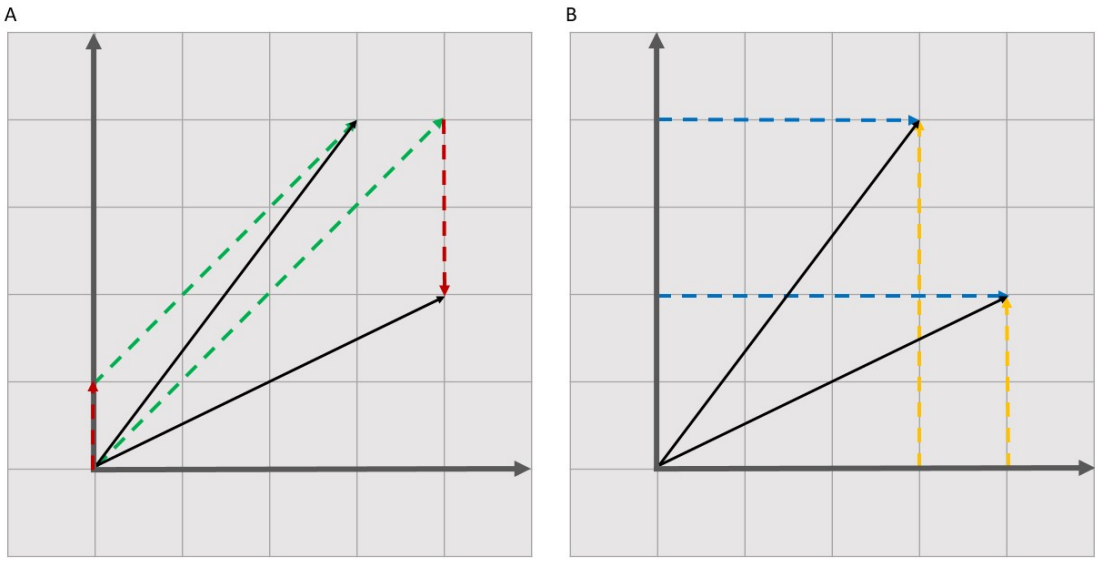


Figure 3.2: A graphical representation of a basis function.
(**A**) The vectors $(3,4)$ and $(4,2)$ are described with the non-orthogonal basis $(1,1)$ (shown in green) and $(0,1)$ (shown in red). (**B**) The vectors $(3,4)$ and $(4,2)$ are described with the orthogonal basis $(1,0)$ (shown in blue) and $(0,1)$ (shown in yellow).

3.3.1 Basis function expansions

In contrast basis function expansions, when taken to a high enough degree, extract all of the information that is encoded within morphology of a cell (Pincus and Theriot, 2007). This is achieved by representing the morphology as a unique mathematically function. The function is then decomposed into a linear combination of functions, known as a basis (Flusser et al., 2009). The individual morphologies can then be differentiated from one another dependent on their specific basis function combination, known as the moments of a function.

An example of a basis can be seen by representing points on a 2D grid. In this example the *space* is all possible vectors of the form (x,y) where x and y are real numbers. The vectors $(1,1)$ and $(0,1)$ then form the basis by which all possible elements, vectors, in the space can be represented as a linear combination. Hence the point $(3,4) = 3(1,1) + 1(0,1)$, and the point $(4,2) = 4(1,1) - 2(0,1)$ (Figure 3.2A). Whilst this example is trivial the same

concept can be extended to higher dimensional vector spaces and into function spaces via a polynomial basis, the set $\{1, x, x^2\}$ is an example of a polynomial basis.

Orthogonal basis

Whilst the basis $(1,1)$ and $(0,1)$ was used in the previous example is does not mean that it is the only eligible basis in the space. On the contrary, there are often multiple bases that can be used within a given space i.e. the basis $(2,1)$ and $(1,2)$ is also valid. However, some bases are used more often because they possess a desirable characteristic, orthogonality.

Formally two functions or vectors are orthogonal if their dot product is $= 0$. Although graphically a pair of vectors can be seen as orthogonal if they intersect at right angles to one another. Thus, the basis $(1,1)$ and $(0,1)$ is not orthogonal as seen by the 45°intersection. In contrast the basis $(1,0)$ and $(0,1)$ is orthogonal. Hence a linear combination of the basis $(1,0)$ and $(0,1)$ can be used to represent the point $(3,4) = 3(1,0) + 4(0,1)$ and the point $(4,2) = 4(1,0) + 2(0,1)$ (Figure 3.2B). The benefit of an orthogonal basis is that the system is more straightforward to solve and then interpret, especially in higher dimensional spaces. An orthogonal basis is also beneficial when calculating the moments of a function as the orthogonality is inherited by the moments. This is important because it means that each additional moment is then independent. Hence the same level of accuracy can be achieved with fewer moments compared to a non-orthogonal basis which has implications on the performance, as discussed in the following section.

Two orthogonal bases that are commonly used in image analysis are Fourier series (Tweedy et al., 2013) and Zernike polynomials (Zernike and Stratton, 1934). Zernike polynomials are often favoured as the x and y components can be evaluated together and the resultant Zernike moments have a low reconstruction error (Teh and Chin, 1988). As a result, Zernike moments have been used previously to evaluate the morphology of cancer cells in still images with great success (Alizadeh et al., 2016; Tahmasbi et al., 2011). However, their application to live cell imaging remains limited. Extending their application to live cell imaging therefore offers an exciting opportunity to quantify the morphological dynamics of individual cells through time, the central theme to this thesis.

3.3.2 Method of moments

Moments are most commonly encountered in statistics to characterise the shape of a distribution with a scalar quantity. For example, if a normal distribution is fitted to a sample of data the distribution will be centred around the average value of the sample. Likewise, the "bell" of the distribution will be more or less diffuse dependent on the spread of the data. These two characteristics are referred to as the mean and variance. Yet, more generally they are the first and second order moments of the function, in this case the function is a normal density function. Moments therefore act as method to capture the features of a sample with respect to a general form.

Whilst only the 1st and 2nd order moments are normally used in statistics higher order moments also exist, the 3rd and 4th order moments are known as skewness and kurtosis respectively. The motivation for using higher order moments is that they extract more detail about the shape of the function. This then reduces the number of samples that the function can represent and allows for greater differentiation between groups. Furthermore, if moments are taken to a high enough order, then every distribution can be represented uniquely. This is an important point in image analysis as it ensures that every shape can be distinctly defined if a high enough moment order is used. Albeit, determining what moment order is *high enough* can be challenging.

Moment error

A moment is defined as a projection of a function onto a polynomial basis. The general form for a moment of order $(p+q)$ in the image plane ξ is:

$$M_{pq} = \iint_{\xi} \underbrace{\psi_{pq}(xy)}_{\substack{\text{Basis}\\ \text{Function}}} \underbrace{f(x,y)}_{\substack{\text{Image}\\ \text{Function}}} dx\,dy \quad : \quad p,q = 0,1,...,\infty \tag{3.1}$$

and if moments are taken to the order $(p+q) = \infty$ than an image will be reconstructed exactly. However, in practice, moments are taken to the order $(p+q) = \nu : \nu << \infty$. This allows the method of moments to be tractable but it also helps to reduce the effect of background

noise. The background noise arises a result of numerical errors that occur during practical applications. Whilst unavoidable the background noise does mean that there is an error between the original image, $f(x,y)$, and the reconstructed image $\hat{f}(x,y)$. Hence minimising the error is essential to ensure that the moments accurately represent the original image (Liao and Pawlak, 1996).

Two factors that heavily influence the size of the error are the max moment order v, which is positively correlated with the error, and the choice of basis. Thus, if reducing the max moment order helps to minimises the effect of background noise. Then choosing a basis that requires a smaller moment order, such as an orthogonal basis, is a advantageous.

Invariance criteria

The choice of basis also effects whether the moments are invariant to translation, rotation, and scaling. Satisfying these 3 invariance criteria is essential to ensure that any difference between moments is solely due to a difference in shape rather than prospective i.e. the same image is recorded regardless of whether it is upside down.

Translational invariance can be achieved by translating each object to have the same centre of mass i.e. the centre of mass is always around the origin, $(0,0)$. Likewise scaling invariance can be achieved by scaling each object by a constant with respect to its average radius (Flusser et al., 2016). Finally, whilst rotational invariance can be achieved with Cartesian coordinates (Hu, 1962) the reconstruction power is often very low. Thus, circular moments are commonly used instead as they are naturally invariant to rotation and can be adopted by transforming the image to a Polar coordinate system.

3.3.3 Zernike moments

A special form of circular moments that keep their magnitude constant under rotation are Zernike moments (Zernike, 1942). The moments are calculated using a Zernike polynomial basis which is orthogonal on the unit disk, $(x^2+y^2) \leq 1$, meaning that each moment is also independent. The combination of rotational invariance and reduced background noise from

the orthogonal basis means that Zernike moments are frequently used in image analysis to quantify objects and shapes (Flusser et al., 2009; Liao, 1994).

Zernike moment definition

Zernike moment are defined in a discrete image as (Shutler and Nixon, 2006):

$$Z_{m,n} = \frac{m+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} P_{xy}[V_{m,n}(x,y)]^* \quad : \quad x^2 + y^2 \leq 1 \tag{3.2}$$

where m defines the moment order and can take integer values from $0 \rightarrow \infty$. The image function is represented by P_{xy} but because an image is measured on a discrete scale, in pixels, the double integral in Equation 3.1 is replaced with a double summation in Equation 3.2. The double summations starts at $(0,0)$ and finishes at (N,N) where N is equal to the size of the image. Likewise the basis function, the Zernike polynomials, corresponded to $V_{m,n}(x,y)$ where * denotes the complex conjugate. Finally n, the integer of rotation, is either positive of negative so long as the following two conditions are satisfied:

$$m - |n| = even \quad and \quad |n| \leq m \tag{3.3}$$

whereby $|n|$ denotes the absolute value of n. Hence an even moment order will have an integer of rotation that is even, or 0, and an odd moment order will have an odd integer of rotation. The integer of rotation is important as it controls the number of oscillations that occur within each Zernike polynomial which in turn controls the amount of flexibility within the polynomial.

Zernike moment error

Due to Zernike polynomials being orthogonal on the unit disk they can also be expressed in polar coordinates such that:

$$V_{m,n}(r,\theta) = R_{m,n}(r) exp(in\theta) \quad where \quad i = \sqrt{-1} \tag{3.4}$$

and $R_{m,n}(r)$ is the radial polynomial

$$R_{m,n}(r) = \sum_{s=0}^{(m-|n|)/2} \underbrace{(-1)^s \frac{(m-s)!}{s!(\frac{m+|n|}{2}-s)!(\frac{m-|n|}{2}-s)!}}_{\text{Polynomial Coefficient}} r^{m-2s} \quad : \quad r \leq 1 \tag{3.5}$$

where m and n represent the moment order and integer of rotation. In polar form it becomes clear that the polynomial coefficient is dominated by factorials in both the numerator and the denominator. Thus the magnitude of the coefficient tends to become large even when m is relatively small. The combination of a large coefficient with a high polynomial order means that the error between the numerical approximation of the integral in Equation 3.1, calculated by the double summation in Equation 3.2, also tends to become large. Hence this error causes the difference between the original image, $f(x,y)$, and the reconstructed image $\hat{f}(x,y)$. Thus, higher order moments have to be used with caution when quantifying the shape of an object to ensure that the information being retrieved outweighs the numerical error of the moment.

Zernike moment invariance

However, the pre-processing necessary to achieve invariance to translation and scaling causes the first two moments, orders 0 and 1, to no longer be informative. Therefore, the moments have to be removed from any further analysis causing an even high moment order to be used as a result.

$$|Z(m)(n)| = |Z(m)(-n)| \tag{3.6}$$

Likewise rotational invariance is achieved by using the magnitude of the moment rather than the moment itself. Yet, Zernike moments have a positive and negative component dependent on whether $n > 0$ or $n < 0$. Thus, because the magnitude of a complex number is equal to the magnitude of it complex conjugate the positive and negative rotations are no longer distinct (Equation 3.6). Therefore, all analysis strictly considers only the non-negative

rotations of each polynomial order to ensure that the information is not being duplicated (Alizadeh et al., 2016; Shutler and Nixon, 2006; Tahmasbi et al., 2011). However, this then means that the amount of information being retrieved at each moment order is also reduced which further prompts the need for higher order moments, albeit at the cost of detecting further background noise.

3.3.4 Optimal number of Zernike moments

Selecting the optimal moment order is an important yet difficult balance to strike. The moment order needs to be high enough to ensure that sufficient information is captured, and thus unique shapes can be differentiated from one another. Yet the order also needs to be low enough to ensure that any surplus background noise is kept to a minimum and thus the performance is maximised.

Mean squared error

The reconstruction performance for a given moment order γ can be evaluated via the mean squared error (MSE). The MSE is calculated as the sum of the squared error between the original image, $f(x,y)$, and the reconstructed image $\hat{f}(x,y)$ (Liao and Pawlak, 1996). Note that the reconstructed image $\hat{f}(x,y)$ for order γ is formed using all moment values $\leq \gamma$. For example, if the moment order $= 10$ then all previous orders up to and including order 10 are used in the reconstruction. The MSE is then calculated for increasing orders to determine the point at which the background noise outweighs the gain in detail.

Mean squared error results

To determine the optimal moment order a frame was chosen at random from the experimentally evolved time-lapse videos. The frame contained a total of 36 cells with a wide variety of different morphologies. The MSE was then calculated for each of the 36 cells starting at a moment order from 2 through to a moment order of 48. The average MSE was then

calculated for each moment order across the 36 cells to produce an average MSE which was then plotted against the moment order (Figure 3.3)

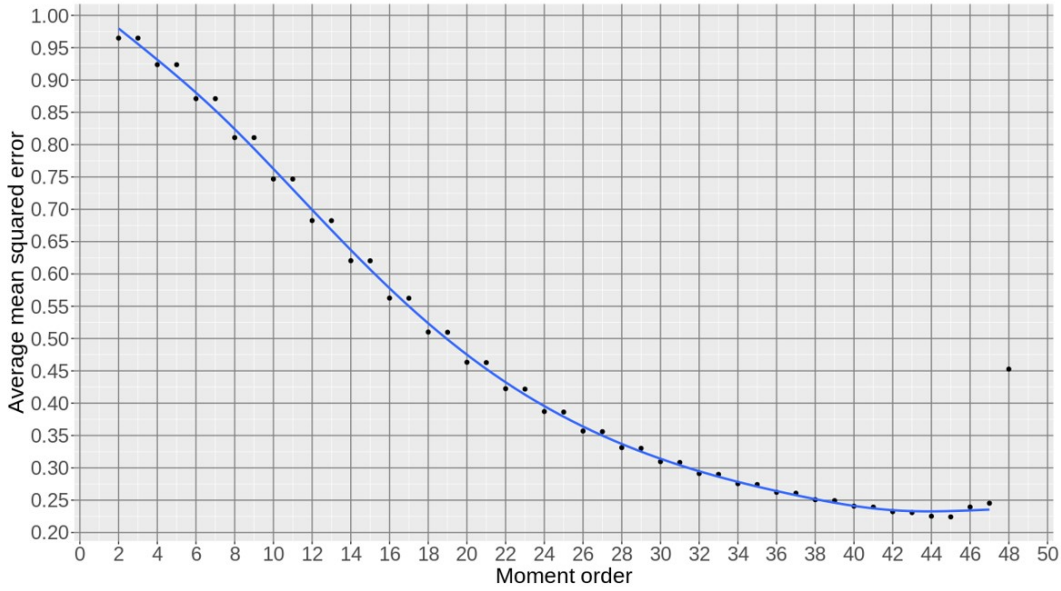


Figure 3.3: A plot of the average mean squared error against moment order.
The average mean squared error (MSE) of reconstruction against the Zernike moment order. The moment order increases from 2 through to 48 in one moment intervals. At a moment order of 2 there is an MSE of 0.965. The MSE then continues to decrease until a moment order of 45 at which the lowest MSE value of 0.224 is obtained. Finally the MSE begins to increase rapidly for moment orders greater than 45 with a value of 0.455 at a moment order of 48. The MSE drops below 0.5 at a moment order of 20 to indicate the minimum order at which an informed reconstruction is achieved.

Initially the average MSE decreases quickly as the moment order increases (Figure 3.3). The reduction in average MSE then continues until an inflection point is reached at a moment order of 45. The inflection point indicates the moment order at which the reconstruction error has been minimised. After reaching the inflection point the average MSE then begins to rise rapidly, indicating that any additional moments would only add surplus noise to the reconstruction (Liao and Pawlak, 1996).

However, whilst the reduction in average MSE continues until a moment order of 45 the rate of reduction is not constant. Instead, the average MSE follows an exponential decay whereby the relative reduction in MSE per additional moment order becomes less closer

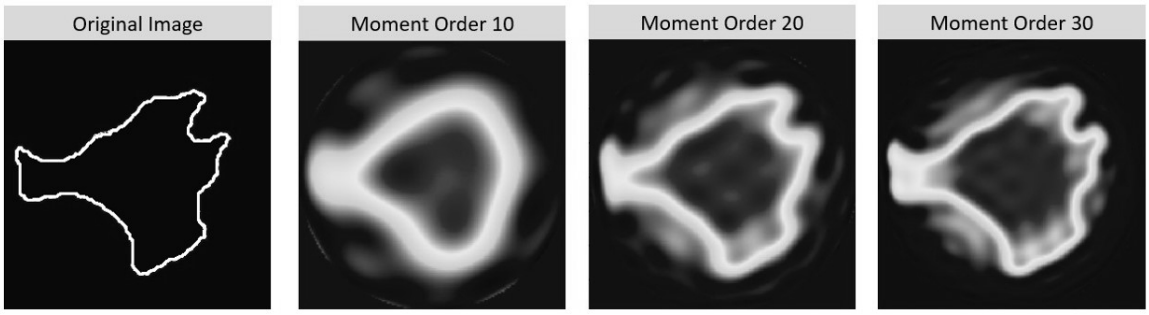


Figure 3.4: The Zernike moment reconstruction performance on cell morphology. The reconstruction performance increases as the Zernike moment order increases. However, the increase in reconstruction performance from a Zernike moment order of 10 to 20 is considerably more compared to the increase in performance from a Zernike moment order of 20 to 30. All of the reconstructions suffer from background noise as seen by the grey pixels that collect around the white outline.

to the point of inflection. For example, a moment order increase from 20 to 30 causes an average reduction in MSE of 0.154. In contrast a similar 10 order moment increase from 30 to 40 causes an average reduction in MSE of 0.069, less than half the reduction for the same 10 order increase (Figure 3.3). This result highlights how higher order moments extract ever finer levels of detail whereas lower order moments capture large structural patterns. Whilst extra detail is useful, especially when categorising morphologies within a still image, in a dynamic setting it can also make it more difficult to determine what aspect of the shape has changed i.e. one large change in shape from a rectangle to a circle or a combination of smaller changes from a smooth to ruffled edge. Hence, the point of inflection can be interpreted as the moment order upper bound rather than a set value.

Another important marker is the moment order at which the MSE drops below 0.5. The MSE is calculated as the sum of the difference in each pixel between the original image, $f(x,y)$, and the reconstruction, $\hat{f}(x,y)$. The two images are both 8-bit grey scale images meaning that each pixel can take a value between 0 and 255, white and black respectively. In the original image, $f(x,y)$, the pixel values are strictly either 0 or 255 as the object shape is known exactly. However in the reconstructed image, $\hat{f}(x,y)$, the pixel values vary between 0 and 255 because the moment order $\gamma << \infty$ (Figure 3.4). Hence if the standardised error of a given pixel is greater than 0.5 the reconstruction of that pixel is no better than chance. Thus

the moment order at which the average MSE drops below 0.5 can be seen as the minimum order for an informed reconstruction of the shape. Across the 36 cells that average MSE first dropped below 0.5 when the moment order $= 20$ (Figure 3.3). Therefore a moment order of 20 can be interpreted as the minimum order needed for an informed reconstruction.

Zernike moment summary

To summarise, the average MSE was calculated to determine the optimal moment order. The lowest reconstruction error is achieved with a moment order of 45. Yet an informed level of reconstruction is achieved with a moment order of 20. Whilst the morphology of each cell needs to be quantified accurately the focus of this thesis is to investigate the change in morphology, which is then used as a proxy for signal processing. Hence the onus is on detecting large morphological changes that demonstrate an active response instead of focusing on the morphology itself. As a result, a moment order of 20 is used for the remainder of this thesis to measure the morphology of each cell, albeit the same qualitative result was also achieved with a moment order of 45.

3.4 Results

3.4.1 Quantifying dispersal in evolved populations

In each frame three metrics were captured for every cell: morphology, spatial location and distance to a neighbouring cell. The morphology was measured with 20 Zernike moments (Zernike, 1942) and then the rate of morphological change was calculated as the Euclidean distance between the vector of moments in frame t and $t+1$ relative to the time between frames (Figure 3.5B). The speed of migration was then calculated as the change in spatial location between consecutive frames (Figure 3.5C). Finally, the distance to the closest neighbouring cell was recorded as the shortest distance from the edge of the cell contour to another neighbouring cell contour without crossing the body of the cell (Figure 3.5D). The

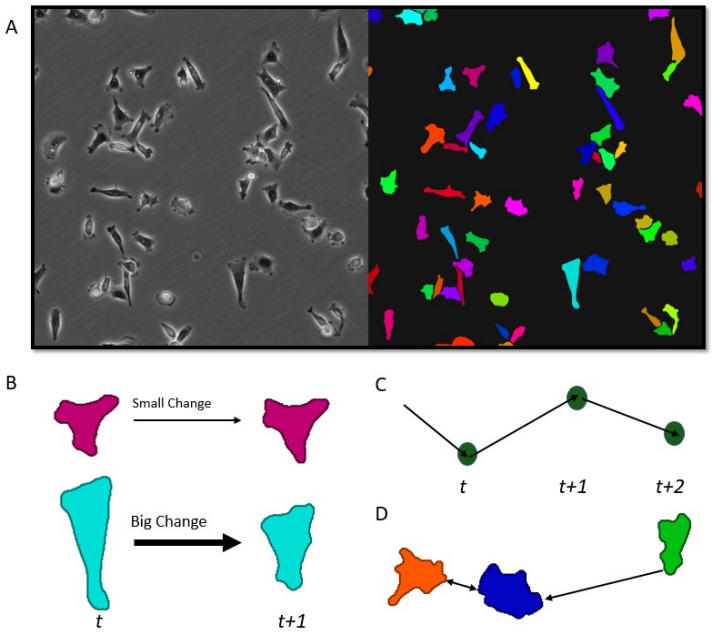


Figure 3.5: Quantifying dispersal from time-lapse videos.
(**A**) Cells were tracked over a 12 hour period with images taken at two minute intervals using phase contrast time-lapse microscopy to generate movies from which morphology could be segmented through the use of a convolutional neural network. (**B**) The rate of morphological change was recorded as the distance between Zernike moments in consecutive frames. (**C**) The speed of migration is calculated as the distance between the spatial location of cells in consecutive frames. (**D**) The distance between neighbouring cells is quantified as the shortest distance between the contour of one cell and the contour of another. The direction of the arrow points from a given cell to the point on the contour of the closest neighbouring cell.

three metrics were then averaged across the entire trajectory of the cell, providing a summary of the dispersal behaviour for each cell.

After the three metrics were calculated the rate of morphological change and the speed of migration were then evaluated to determine whether there was a significant difference among the four populations. An analysis of variance (ANOVA) was used to compare the mean rate of morphological change and the mean speed of migration across all populations, differences in wells were accounted for as a random effect. There was significant variation among the populations in their mean rate of morphological change ($p = 0.0296$, $N = 813$). A post-hoc Bonferroni multiple comparison test was then used to identify which populations were significantly different. The escape populations were found to have a significantly higher

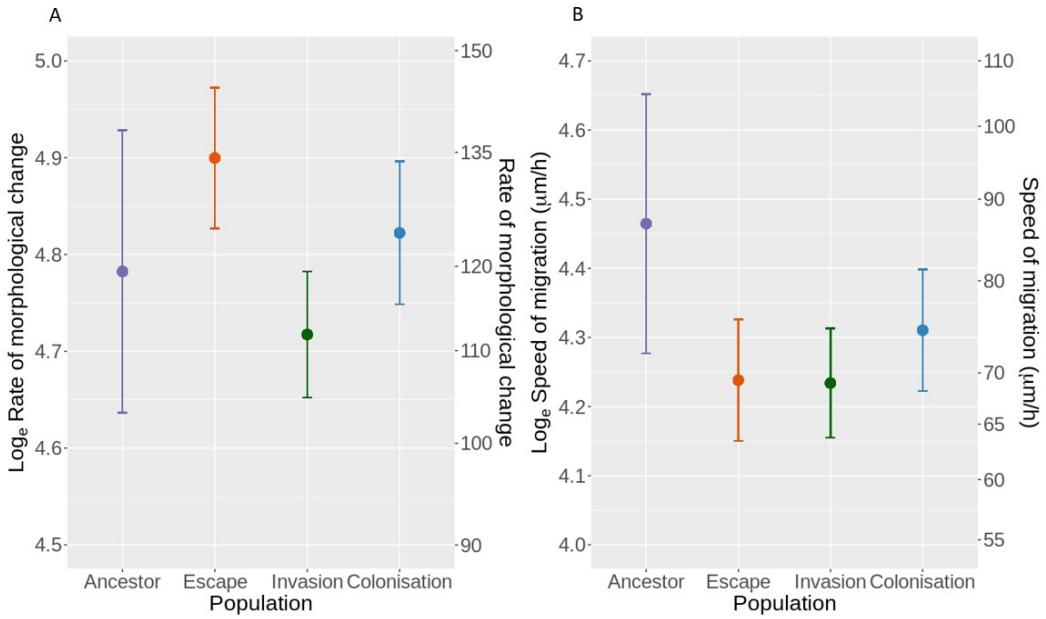


Figure 3.6: Comparing the mean rate of morphological change and speed of migration among the four populations. (**A**) A plot of the natural log-transformed rate of morphological change for each of the four populations. The centre dot signifies the mean rate of morphological change with errors bars signifying 95% confidence intervals. The escape populations had a significantly faster rate of morphological change compared with the invasion populations, p = 0.0152 (N = 813). (**B**) A plot of the natural log-transformed speed of migration for each of the four populations. The centre dot signifies the mean speed of migration with errors bars signifying 95% confidence intervals. There was no significant difference in the average speed of migration among the 4 populations. The mean, standard error and number of observations for each population can be found in Table B.1, Appendix B

rate of morphological change compared with the invasion populations (p = 0.0152, N = 813; Figure 3.6A). There was no significant difference in the mean speed of migration among the four populations (Figure 3.6B).

3.4.2 Speed of migration predicts rate of cell-morphological change in evolved populations

Next, the morphological behaviour was investigated in response to the speed of migration and the nearest neighbour distance. A linear mixed model was fitted across all of the data such that the rate of morphological change was dependent on the speed of migration, the

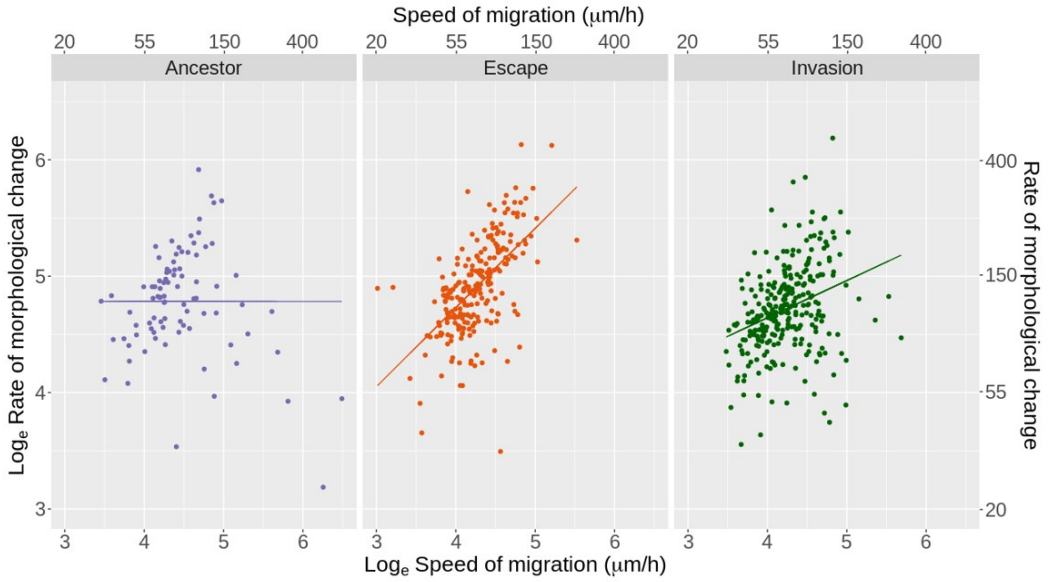


Figure 3.7: The rate of morphological change against the speed of migration.
The natural log-transformed rate of morphological change plotted against the natural log-transformed speed of migration. The straight lines represent the reduced model for each population using only parameters that are significant at the 5% level. The ancestor populations have an intercept-only model fitted (N = 88). The speed of migration is the only significant variable in the escape (N = 230, $p = 1.765 \times 10^{-3}$) and invasion (N = 283, p = 0.018) populations. For both escape and invasion populations the rate of morphological change is positively correlated with the speed of migration, the faster the speed of migration the higher the rate of morphological change.

distance to the nearest neighbouring cell and the interaction of the two (Equation 3.7). The model parameters were selected through a process of forward selection and only included if they were significant at the 5% level. The populations were also included as a fixed effect allowing the intercepts and slopes to vary between populations. The significant parameters were then used to fit a reduced model to the ancestor, escape and invasion populations (Figure 3.7).

$$\begin{matrix}\text{Rate of}\\ \text{morphological}\\ \text{change}\end{matrix} = \alpha + \beta_1 * \left(\begin{matrix}\text{Speed of}\\ \text{migration}\end{matrix}\right) + \beta_2 * \left(\begin{matrix}\text{Distance to}\\ \text{nearest neighbour}\end{matrix}\right) + \beta_3 * \left(\begin{matrix}\text{Speed of}\\ \text{migration}\end{matrix} * \begin{matrix}\text{Distance to}\\ \text{nearest neighbour}\end{matrix}\right) \quad (3.7)$$

In the ancestor populations neither the speed of migration nor the distance to neighbouring cells significantly affected the rate of morphological change. As such, an intercept only model was fitted to the data (Figure 3.7). However, the intercept model explained only a small proportion of the variance, (marginal $R^2 = 0$) (Nakagawa and Schielzeth, 2013). This might therefore suggest that the rate of morphological change is either highly stochastic, or that it depends on factors not included in the model.

In contrast, the speed of migration was significant in both the invasion and escape populations such that it was positively correlated with rate of morphological change, (β = 0.680 and 0.319 respectively: Figure 3.7). Furthermore, the escape and invasion models also both explained a significant proportion of the variation (marginal $R^2 = 0.347$ and 0.099 respectively). Finally, to ensure that the results were not affected by a small cluster of potential outliers the same analysis was repeated after having removed any influential data points (Figure B.2, Appendix B), defined by a Cook's distance > (4 / N) where N is the sample size (Bollen and Jackman, 1985). Yet, the same qualitative relationship was still present.

The steeper slope in the escape populations compared with the invasion populations might therefore suggest that selection for escape favours cells that can change their morphology rapidly when migrating at higher speeds. This might be a result of the collagen escape assay being a 3D to 3D environment compared with the 2D to 3D environment of the Matrigel invasion assay. However, this could also be due to the different number of rounds of selection between the two assays, or a difference in the strength of selection within each.

3.4.3 Spatial density affects morphological dynamics

Finally, the colonisation populations displayed a more complex morphological behaviour dependent on the speed of migration, the distance to the nearest neighbouring cell and the interaction of the two: as the distance between neighbouring cells increases, the relationship between the rate of morphological change and the speed of migration becomes negative (Figure 3.8A). Hence when close to a neighbouring cell, the rate of morphological change is positively correlated with the speed of migration: a faster speed of migration results in a

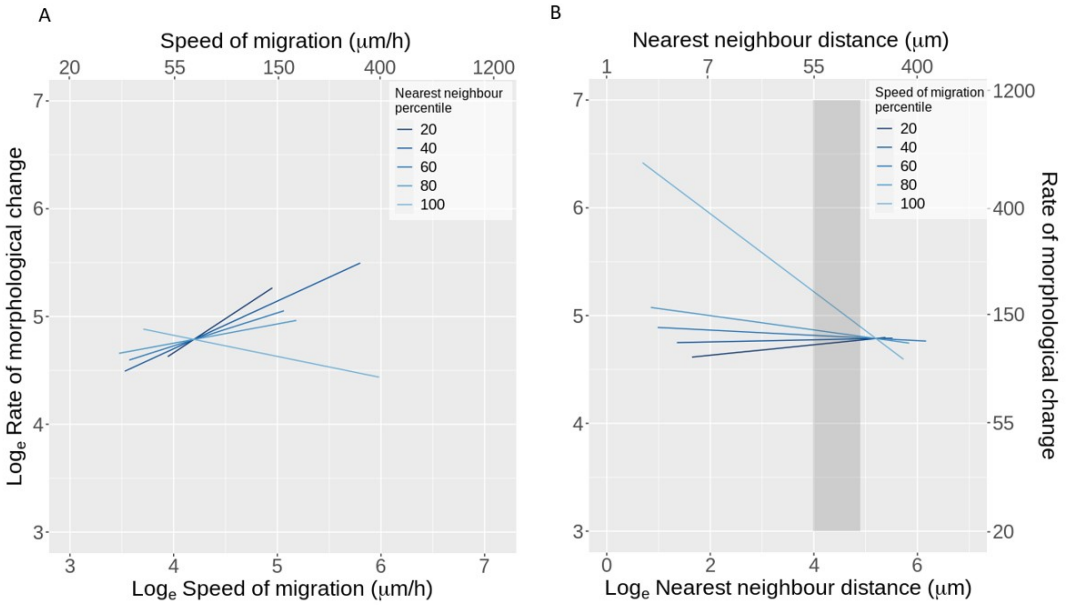


Figure 3.8: A dynamic switch in the morphological behaviour within cells selected for colonisation. Data points have been removed to highlight the behaviour of the model, the same model with data points can be seen in Figure B.1, Appendix B. The speed of migration ($p = 5.418 \times 10^{-14}$), the distance to the nearest neighbouring cell ($p = 2.207 \times 10^{-10}$) and the interaction of the two ($p = 2.219 \times 10^{-11}$) was significant in the colonisation populations (N = 212). (**A**) The predicted natural log-transformed rate of morphological change against the natural log-transformed speed of migration. The shaded lines indicate the natural log transformed nearest neighbour percentile. The lighter the line, the further away from a neighbouring cell with distance values ranging from 2μm - 477μm. (**B**) The predicted natural log-transformed rate of morphological change against the natural log-transformed nearest neighbour distance. The shaded lines indicate the speed of migration percentile. The lighter the line the faster the speed of migration. The shaded region indicates the range of distances over which there is no significant relationship in the rate of morphological change and the speed of migration when the data is centred at these distances, between 57.9μm and 147.2μm.

higher rate of morphological change. However, when the distance between neighbouring cells is large and a cell is isolated, the rate of morphological change is negatively correlated with the speed of migration: a faster speed of migration has a lower rate of morphological change. Furthermore, the same analysis was also repeated after the removal of any potentially influential data points and the interaction term was still significant in the colonisation populations (Figure B.2, Appendix B). Finally, the colonisation model also explained a

significant proportion of the variation in the rate of morphological change (marginal R^2 = 0.236).

The switch in morphological behaviour was then examined further to determine whether the change in behaviour was gradual or sudden. This hypothesis was investigated by centring the nearest neighbour data a distance x and then refitting the same morphological change model (Equation 3.7). The speed of migration was then evaluated to determine whether it was still significant in the model. If the speed of migration was not significant at a distance x then there was no significant difference in the rate of morphological change for cells migrating at different speeds. The method was then repeated for different values of x to find a range of distances over which the speed of migration was not significant. The smaller the range the more sudden the switch.

At nearest neighbour distances between 57.9μm and 147.2μm the speed of migration was not significant in the model, as seen by the shaded region in Figure 3.8B. Therefore, at distances < 57.9μm or > 147.2μm the speed of migration is significantly related to the rate of morphological change. The small range of distance values suggests that the cells have a high degree of sensitivity to the location of neighbouring cells. Interestingly, the range of distance values coincides with values from the literature whereby cells within a tumour core have been seen to display a correlated mode of migration at spatial distances < 50μm compared with distances greater than 250μm (Staneva et al., 2019).

3.5 Discussion

Novel phenotypic analysis was conducted across 4 experimentally evolved populations of MDA-MB-231 breast cancer cells to investigate their morphological behaviour during dispersal. Combining experimental evolution with computer vision enabled a multidimensional data set to be formed that captures the dispersal dynamics of individual cells within each population. This data set has then been used to build a data driven morphological model that has uncovered fundamental dynamics at a cellular level and is capable of distinguishing cells selected for colonisation. Due to its unique and powerful nature this data set is further used

in Chapters 4 and 5 to investigate other dispersal characteristics that may influence metastatic success.

3.5.1 Navigating through a complex environment

The continuous flow of cells through the microenvironment creates a landscape that is both spatially and temporally heterogeneous (Yuan, 2016). This landscape variability might explain the correlation between the rate of morphological change and the speed of migration for both the escape and invasion populations (Figure 3.7). The collagen escape and Matrigel invasion assays used to select the escape and invasion populations are both porous and complex (Anguiano et al., 2017), but yet they are also malleable. This malleability means that large structural changes can occur within the environments and migration routes that were previously accessible may become blocked. Hence a cell may need to respond to the environment changes by also changing its own morphology to ensuring that it can continue to migrate and does not become trapped. Likewise, as the speed of migration increases, an increase in the rate of morphological change might also be necessary to ensure that the cells are not temporarily stuck by any potential obstacles. This would then also explain why there is no correlation in the ancestor populations where the environment remains constant and thus would be no selective advantage to this behaviour.

In addition to comparing the dispersal behaviour of the final populations it would also be valuable to examine the intermediate populations, the cells that were selected after the first or second round of selection. If the intermediate populations displayed the same behaviour as the final population, then this might suggest that a single sub-population was selected after the first round. However, it may also be that the intermediate populations display a more gradual change in behaviour whereby the steepness of the slope increases with each round of selection. This would therefore suggest that the number of sub-populations are being iteratively reduced with each round of selection. In either case it would give an insight into the clonal heterogeneity of the original ancestor population (McGranahan and Swanton, 2017) as well as comparing the strength of selection between in both the escape and invasion assays.

3.5.2 Forging a path within a crowd

Distant-site colonisation requires a cell to switch from a mode of long-range dispersal and focus on re-initiating aggressive proliferation; the subsequent increase in local cell density may reduce available space and thus intensify competition. A similar selective pressure can be seen in the colonisation assays. In contrast to the ancestor, escape and invasion populations, where cells are periodically moved to a new expansive environment, the colonisation population remain fixed. As such in addition to the structural changes that occurred within the microenvironment there is also a high density of cells migrating locally as thus the cells themselves could block potential migration routes, therefore explaining the significance of the neighbour location in the model. This hypothesis would also explain the interaction that is observed between neighbouring cells. If a cell is migrating at a high speed and is close to other neighbouring cells, then changing its morphology rapidly might be necessary to avoid other cells that are changing location dynamically. However, when isolated the location of neighbouring cells is no longer of concern and thus a reduction in the rate of morphological change might allow a cell to conserve vital resources.

The significance of the neighbour sensitivity may also suggest that the ability of a cell to sense contact has been re-acquired within the colonisation population. A loss of contact inhibition is seen as one of the earliest developments in cancer progression as it allows aggressive proliferation to ensue, which in turn gives rise to the formation of a primary tumour (Pavel et al., 2018). However, the high degree of neighbour sensitivity seen in Figure 3.8 questions whether contact sensing is in fact lost, or instead down-regulated earlier in the metastatic cascade. If true, this could suggest that cells selected for distant-site colonisation are able to vary their own contact sensing ability dependent on the exogenous environmental stresses they encounter.

3.5.3 Detecting complex phenotypic behaviours

In summary, interpreting cellular morphology as a dynamic process provides novel insight into the behaviour of breast cancer cells, and furthers the understanding of the phenotypic

route to metastasis. A future next step will be to evaluate the morphological dynamics in a native 3D environment (Petrie and Yamada, 2012) and in the vicinity of stromal cells such as fibroblasts which are known to have a critical role in metastasis (Malanchi et al., 2012). The presence of stromal cells might also change the relationship seen within our escape and invasion populations, as cells would then be able to interact via matrix metalloproteinases. Thus, rather than needing to change their morphology quickly to prevent being trapped, they could exploit the matrix metalloproteinases to cut themselves free, as seen previously during metastatic dispersal (Page-McCaw et al., 2007). Likewise, the behaviour seen within the colonisation populations may also change when the neighbouring cell is non-cancerous. If so, then it may suggest that cancer cells selected for colonisation are actively registering one another rather than simply avoiding another obstacle. Nevertheless, this work highlights the power of phenotypic analysis in discovering complex emergent behaviours that would not have been apparent from genetic data.

Chapter 4

Heterogeneity in cancer cell signal processing

4.1 Introduction

Heterogeneity is widespread in nearly all types of cancer (McGranahan and Swanton, 2017). Genetic variation can be seen between patients with the same cancer type, intertumoural heterogeneity, as well as between individual cells within a given tumour, intratumoural heterogeneity (ITH) (Burrell et al., 2013; Dagogo-Jack and Shaw, 2018). Clinically the degree of ITH is important, as an elevated level of ITH is seen to correlate with an increased likelihood of therapy resistance and worse patient outcomes (Morris et al., 2016). As a result, deciphering the evolutionary path that leads to increased tumour diversity is essential in trying to predict, and ultimately constrain, cancer progression (Turajlic et al., 2019).

Genetic heterogeneity and therapy resistance

High levels of ITH are expected to increase the rate of tumour evolution, and thus the likelihood of therapy resistance, through two different mechanisms. Firstly, tumour evolution is driven, in part, by mutations that randomly appear within the population. The majority of these mutations either have either a negligible or negative effect on the individual's fitness. However, occasionally, a beneficial mutation will appear that increases the individual's

fitness and is thus selected (Martincorena et al., 2017). A high degree of genetic variability is therefore expected to increase the breadth of mutations that appear and thus the likelihood of a beneficial mutation emerging (McGranahan and Swanton, 2015). Secondly, an increase in genetic diversity means that relative to normal tissue there is a reduction in cellular homogeneity, and thus a corresponding increase in cellular competition. This aggressive competition then promotes the outgrowth of mutant clones (Parker et al., 2020), which further increases the rate of tumour evolution and disease progression (Section 1.1.2).

4.1.1 Phenotypic heterogeneity

Whilst extensive genetic heterogeneity is common, linking specific genetic changes to functionally different phenotypic traits remains an enigma (Graham and Sottoriva, 2017). Resolving this shortfall however is important because adaption occurs at a phenotypic level. Thus, without an accurate genotype-to-phenotype mapping, it is hard to differentiate the key genetic changes that are critical for cancer progression from the remaining background noise (Yi et al., 2017).

In addition, cancer also has a high degree of plasticity (Sharma et al., 2010). Non-genetic factors such as changes in gene expression have a large, but transient effect on the phenotype of a cell (Burrell et al., 2013; Marusyk and Polyak, 2010; Meyer and Heiser, 2019). This causes phenotypic variation to exist on both a cellular and evolutionary timescale. Yet, due to logistical challenges, both levels of variability are rarely evaluated together and instead the traits are assumed to be constant at the level of the cell. For example, the migration behaviour of a cell is typically summarised by a single quantitative value over the entire migration trajectory. Whilst this assumption can provide valuable insight (Chapter 3), it also relies upon the local microenvironment remaining equally constant. That is, the same environmental signals are received by a given cell over its lifespan. However, this assumption is often not true, especially during metastasis.

Short term phenotypic variability in metastasis

Firstly, the local microenvironment is known to have a high degree of spatial and temporal variability (Yuan, 2016). Hence the environmental signals that an individual cell will receive are inherently time dependent. Secondly, each individual cell is moving. Thus, even if the surrounding environment remains temporally constant, an individual cell will still experience a change in environmental signals as a result of its own migration. A high degree of short term phenotypic variability may therefore be advantageous during metastasis to compensate for the variability within the surrounding environment. As a result, cells selected for colonisation at a distant site are expected to change their phenotypic behaviour more frequently in response to subtle changes in environment conditions.

Similarly, once the temporal variation has been partitioned, the degree of within population variability is also expected to be higher within cells that have been selected for distant site colonisation. This expectation is based on the assumption that if a high degree of temporal variation is advantageous, then all cells selected for distant site colonisation will be highly flexible. In turn, the high degree of short term phenotypic variability acts as a buffer against environmental changes, akin to dispersing in the first place (Bowler and Benton, 2005). Yet, if all cells are highly responsive to temporal changes, then the environment is effectively constant through time. Therefore, ecological dispersal theory would predict that multiple stable solutions can exist at an evolutionary optima (McPeek and Holt, 1992). Hence the degree of heterogeneity within the population is also be expected to be higher.

Chapter overview

The two hypotheses that cells selected for distant site colonisation have a higher degree of phenotypic variability and heterogeneity can be tested by evaluating the signal processing behaviour through time within each of the experimentally evolved population of cells (Section 3.2.1). The chapter begins by first explaining how the degree of short term phenotypic variability can be quantified within each population by leveraging the individual time series data that is collected for each cell during tracking. The advantages of using time series migration data are then explored as well as certain issues that need to be considered during

the model building process e.g. how to account for different sources of variability. A temporally dependent morphological model is then built that also includes the influence of known covariates such as the speed of migration and the nearest neighbour distance. The model is then fitted to the time series data at a population level, before then being refitted at a single cell level. Finally, the results are discussed to compare the different levels of short term phenotypic variability and heterogeneity within each experimental population.

4.2 Time series data

Single cell tracking records the position of a cell as a point, or a collection of points, repeatedly over a set period of time. The migratory record can then be used to measure individual cellular traits such as the rate of morphological change or the speed of migration. The measured traits can be used to compare different cellular populations or to build quantitative models that characterise complex cellular behaviours.

Typically, the first stage of analysis involves calculating the average value of a trait over the entire trajectory of the cell. This means that for a given trait each cell is represented by a single datum. Whilst effective, this also assumes that the temporal variation in the trait remains constant over the migration trajectory. Yet in practice this is often not true. Traits such as cell speed are known to change frequently during migration however this variability is rarely captured (Figure 4.1). As a result, the migratory dynamics within a population are summarised as a combination of fixed behaviours.

Modelling temporal data

An alternative approach is to leverage the temporal structure within the data and to acknowledge that the value of a trait at time t may depend on the value of the trait at time $t-1$. Hence the observed time series is then assumed to be generated by an underlying process, such as a random walk, where each observation is a random variable. As a result, each individual trait is then represented by a series of data points rather than a single datum. This means that each trait can then be modelled temporally and at a single cell level. However,

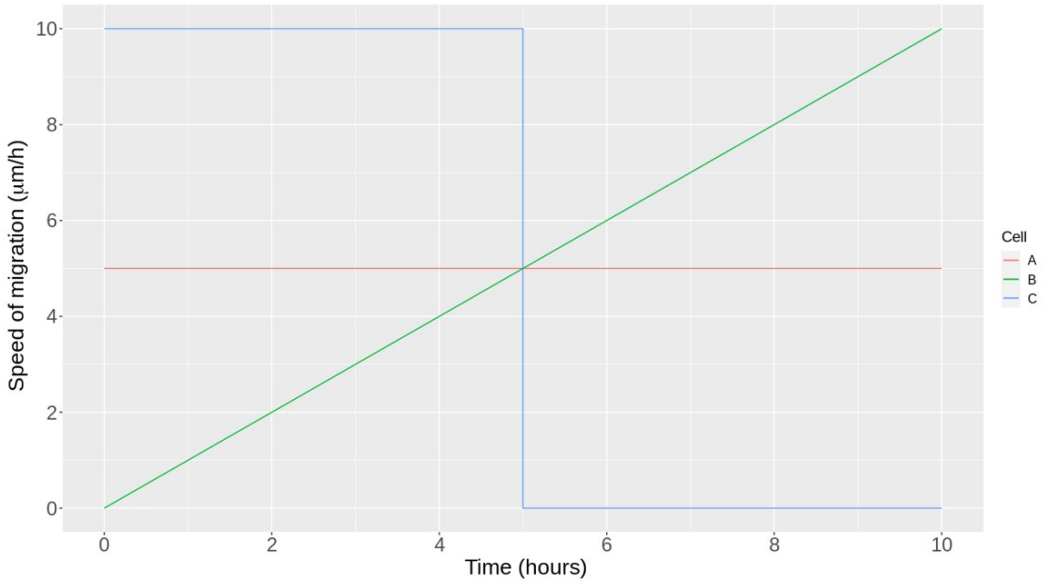


Figure 4.1: A simulated example of the migration behaviour in three cells that have the same average speed of migration. A plot of the migration speed for 3 simulated cells over a 10 hour time period. Cell A migrates at a constant speed of $5\,\mu m/h$. Cell B begins from a static position and then increases its speed of migration linearly over the 10 hour time period reaching a max speed of $10\,\mu m/h$. Cell C migrates at a speed of $10\,\mu m/h$ for the first 5 hours before then stopping and remaining static for the last 5 hours. Whilst the 3 different behaviours are distinct, they each migrate a distance of $50\,\mu m$ over a 10 hour time period and therefore have the average speed of $5\,\mu m/h$.

the temporal structure does also add additional challenges to the model building process. The dependency between consecutive data points, known as autocorrelation, is problematic because it violates an underlying assumption of linear regression. If not resolved the presence of autocorrelation can cause an underestimation of parameter errors which many in turn lead to incorrect conclusions being drawn.

4.2.1 Autoregressive-moving-average model

Fortunately, a wide variety of approaches exist to account for the presence of autocorrelation and therefore model time series data. One of the most common methods is to model the time series as a combination of two polynomials, known as an autoregressive-moving-average model (ARMA). The autoregressive polynomial (AR) models the current value at time t as a

linear combination of historic values plus some stochastic error. Then the moving average polynomial (MA) models the current value at time t as a linear combination of the historic errors plus some stochastic error. In both models the number of historic values and errors are set in advance as hyper parameters whilst the parameters are estimated from the data. The combination of low model complexity and efficient parameter computation has meant that ARMA models are extremely popular in time series analysis, especially when there is an onus on predicting future values as would be the case in a financial setting (Ltkepohl, 2007).

Disadvantages of ARMA modelling

Nevertheless, ARMA models do have limitations, many of which may explain why they are so rarely used when analysing cell migration data. Firstly, ARMA models approximate the underlying dynamics rather than trying to understand them directly. Whilst this enables a prediction to be made, gaining an insight into the underlying dynamics is often the main objective within a biological setting. Secondly, ARMA models do not naturally handle irregular or missing data. This can be problematic in the context of cell migration as data is rarely complete. Missing values can occur due to inaccuracies in the tracking process, or due to logistical issues such as a cell partly leaving the field of view or being excluded because of a division event (Section 3.2.2). Finally, ARMA models are primarily a univariate modelling technique, they model a single time series. Yet in cell tracking experiments a time series is collected for each individual cell. Hence the data is both cross sectional and temporal, known as longitudinal data. The objective therefore is to understand the processes that are generating the time series as well as being able to contrast between the different cells. As a result, a more flexible approach is needed when modelling cell migration data known as state space modelling.

4.3 State space modelling

State space modelling (SSM) is a form of hierarchical modelling that partitions the variation within a time series into two separate models: state and observation (Figure 4.2). The state

model is designed to reflect the true underlying dynamics of the system. The system is assumed to have a temporal structure, whereby the value at time t is dependent on the value at time $t-1$, and evolve through time as a stochastic process. However, the state is also a latent process that cannot be observed directly. Thus, the observation model functions as a link between the state dynamics and observed time series whilst also accounting for the variability introduced by the sampling procedure. The observations are then assumed to be independent once the temporal structure has been accounted for by the state model (Durbin and Koopman, 2001).

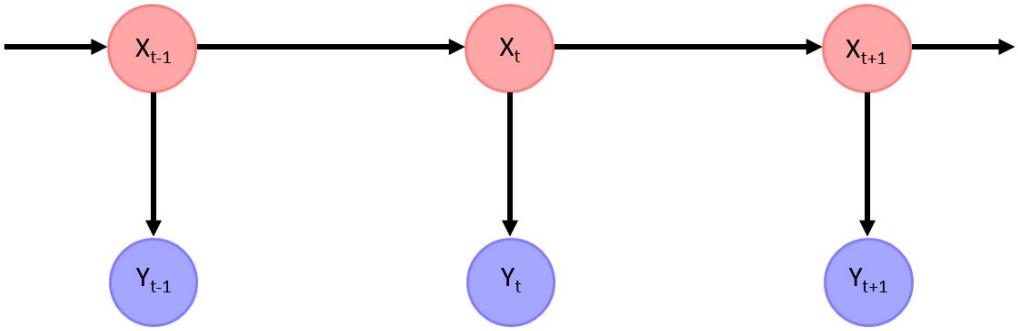


Figure 4.2: The structure and progression of a univariate state space model.
The underlying states, in red, have a temporal structure where the state at time t, $\mathbf{x}_t$, is dependent on the state at time $t-1$, $\mathbf{x}_{t-1}$. The observations, in blue, are then related to the states via the observation equation and are assumed to be independent once the temporal structure has been accounted for by the states.

An example of an SSM can be seen by modelling the change in water depth of a lake. The true depth will change day - to - day in accordance to ratio of rain fall and evaporation. Hence this process has a temporal structure where the depth of the lake on a given day is a function of the previous day plus some stochastic variability. These dynamics are then characterised by the state equation. However, it is not possible to obtain an exact measure of

the water depth due to the uneven surface on the bottom of the lake which also changes day - to - day. Thus, the observation model accounts for the experimental variability that links the underlying dynamics with the observed data.

The ability to partition different sources of variability has meant that SSMs are often used in an ecological setting where data is notoriously difficult to obtain and often has a high degree of experimental variability (Aeberhard et al., 2018). However, their application in modelling cell migration data has been far more limited (Svensson et al., 2018). Extending the use of state space modelling to single cell migration data therefore offers an exciting opportunity to test novel hypotheses and gain insight into previously unknown biological phenomena.

4.3.1 Model structure

The structure of a time invariant SSM with Gaussian errors is defined as (Holmes et al., 2012):

$$\mathbf{x}_t = \mathbf{B}\mathbf{x}_{t-1} + \mathbf{u} + \mathbf{C}\mathbf{c}_t + \mathbf{w}_t \quad \textit{where} \quad \mathbf{w}_t \sim \text{MVN}(0, \mathbf{Q}) \tag{4.1}$$

$$\mathbf{y}_t = \mathbf{Z}\mathbf{x}_t + \mathbf{a} + \mathbf{D}\mathbf{d}_t + \mathbf{v}_t \quad \textit{where} \quad \mathbf{v}_t \sim \text{MVN}(0, \mathbf{R}) \tag{4.2}$$

where Equation 4.1 represents the state model and Equation 4.2 represents the observation model. The state of the system at time t is represented by $\mathbf{x}_t$, an $m \times 1$ vector where $m =$ the number of hidden states in the model. Likewise, the observations at time t are represented by $\mathbf{y}_t$, an $n \times 1$ vector where $n =$ the number of observed time series. The number of states, m, is often equal to the number of observed time series n. However, in some cases where n is large, there maybe a belief that multiple observations arise from the same generative process and thus $m < n$. The ability to include prior beliefs into the model structure is a powerful characteristic of SSMs and it features across multiple aspects of the model structure.

Matrix mapping

The current value of the state, $\mathbf{x}_t$, is related to the previous state value, $\mathbf{x}_{t-1}$, via the state matrix, $\mathbf{B}$. The state matrix is an $m \times m$ matrix where the diagonal elements represent the interaction of the state with itself, and the off diagonal elements represent the inter-state interactions. The ability to estimate the effect of inter-state interactions could be particularly useful in modelling competition assays where an increase in the size of population A causes a corresponding decrease in the size of population B. However, for the remainder of this thesis no inter-state interactions are considered and thus the state matrix remains strictly diagonal. If the state matrix is equal to the identity matrix than the system progresses as a random walk.

Similarly, the observation value at time t, $\mathbf{y}_t$, is related to the state value at time t, $\mathbf{x}_t$, via the $n \times m$ observation matrix, $\mathbf{Z}$. The structure of the matrix $\mathbf{Z}$ is important as akin to the design matrix in a linear regression context, it relates the underlying generative processes to the observed time series. If each cell is believed to originate from a unique underlying process, then the structure of $\mathbf{Z}$ will be equal to the identity matrix. However, cancer develops as a clonal process. Thus, the behaviour of multiple cells maybe expected to arise from the same underlying generative process. Hence the total number of underlying processes may be considerably less than the number of observed time series, $m << n$.

One option is to have an underlying process for each population and then estimate the effect size for a given process on an individual cellular basis. However, this structure assumes that each population specific process occupies a distinct phenotypic subspace. This assumption could be true for the experimentally evolved populations that experience independent selective pressures (Section 3.2.1). Yet, in vivo, each stage of the cascade is completed sequentially. Therefore colonisation of a distant site also requires the completion of earlier stages in the metastatic cascade. An alternative option is to relax the condition of independence and to assume that each underlying process effects the behaviour of every cell. Thus the behaviour of a given cell is represented as a linear combination where the effect size of a given process is estimated from the data. This approach is known as dynamic factor

analysis and has been adopted for the remainder of this chapter (Harvey, 1990; Zuur et al., 2003).

Weight vectors

The mean state vector, $\mathbf{u}$, and mean observation vector, $\mathbf{a}$, are $m \times 1$ and $n \times 1$ vectors that weight the state and process models respectively. In a multivariate system the weighting vectors allow the mean of each state or observation to be independent at equilibrium. Whilst the extra freedom can be useful it also adds an additional overhead to the estimation process. Thus, if the individual equilibrium values are not of interest, the data can be standardised removing the weight vectors from the estimation process. This approach has been adopted throughout this thesis. It is important to stress that the standardisation occurs at an individual level. Thus, the subsequent covariate effects are estimated relative to the individual average not the population average as would be expected in a regression context.

Covariate effects

The flexible nature of SSMs also extends to their handling of covariates that can appear in either the state, $\mathbf{c}_t$, or observation, $\mathbf{d}_t$, models and have dimensions $p \times 1 : m$ and $q \times 1 : n$ respectively. Note that p and q signify the number of covariates in either the observation or state model. Then $1 : m$ or $1 : n$ signify the number of covariate data sources e.g. an individual data stream per state would have dimensions $p \times m$ where as a single data stream would be $p \times 1$. Likewise, whilst the covariates in Equations 4.1 and 4.2 have the subscript t to signify that the input is time varying. The vector can in fact contain a mixture of time varying and time invariant data sources. This freedom is especially useful when including geographical information that is often temporally constant e.g. the source of a chemotactic gradient.

The effect size of the covariates in either the state or process model is defined by the matrices $\mathbf{C}$ or $\mathbf{D}$ with dimensions $m \times p$ and $n \times q$ respectively. Similar to the observation matrix, $\mathbf{Z}$, the structure of the covariate matrices can be defined to allow for effects at different levels in the model. For example, the effect of resource levels on cellular uptake could be

estimated as a constant across all of the cells within the population. That is, the level of uptake is proportional to the resource level for all cells in the population. However, the minimal level of resources needed for a cell to stop migrating maybe estimated on a cell specific basis.

Variance-covariance structures

The characteristic feature of an SSM is the ability to separate different levels of variability between the state and observation models. The variation in the state model at time t is represented by a $m \times 1$ vector $\mathbf{w}_t$, and the observation variation at t by a $n \times 1$ vector $\mathbf{v}_t$. The elements of $\mathbf{w}_t$ and $\mathbf{v}_t$ are assumed to be drawn from separate $m \times m$ and $n \times n$ multivariate normal distributions with mean 0 and variance $\mathbf{Q}$ and $\mathbf{R}$ respectively.

However, in contrast to a linear regression model where the variance, σ, is constant along each dimension of the multivariate normal. The structure of $\mathbf{Q}$ and $\mathbf{R}$ are not restricted. As a result, each dimension of $\mathbf{Q}$ and $\mathbf{R}$ can have the same variance, its own variance, or a mixture of the two. Furthermore, each dimension does not need to be independent. A covariance between dimensions can be estimated in the off-diagonal elements of $\mathbf{Q}$ and $\mathbf{R}$. The versatile variance structure helps to improve the model fit but it also has important implications on the subsequent model interpretation.

An example of how different variance structures effect subsequent interpretations can be seen by modelling the change in population size on the African continent. The model could be specified with a single state variance q which would mean that the population size of each country is temporally independent, and the same across the whole continent. However, changes in population size may also vary differently on a country-by-country basis. As such an alternative model maybe used where separate state variances q_i are estimated for each individual country. Likewise, the assumption of temporal independence between countries may also be invalid. Instead, fluctuations in population size may coincide for adjacent countries. Hence a covariance effect can be estimated between neighbouring countries.

Leveraging the flexible variance-covariance structure of an SSM is also an appealing prospect when modelling single cell migration data. For example, the variance of the

observation model, **R**, can be structured to account for the experimental variability between each well or time-lapse video. Then the state model variance, **Q**, could be structured to account for increases and decreases in the migration speed of neighbouring cells.

Model considerations

Whilst an SSM can have a wide variety of different structures it is important to ensure that the model still reflects the underlying biology of the system (Auger-Méthé et al., 2016). For example, estimating a state matrix, **B**, with inter-state estimates between cells in neighbouring wells would be clearly flawed. Likewise, careful considerations also need to be made for the covariate and variance-covariance structures.

Finally, SSMs are often computationally very expensive to fit. Hence when combined with large multidimensional data sets, such as those generated during cell biology experiments, certain model structures can quickly become intractable (Thygesen et al., 2017). Similarly, SSMs can also suffer from ill-conditioned variance-covariance matrices when the model structure is too complex for the data. This then causes the model to be fitted incorrectly and if not detected can lead to incorrect conclusions being drawn (Auger-Méthé et al., 2016). The implications of an incorrect variance-covariance structure can be seen by the difference between a state variance q that is ≈ 0 and a state variance $= 0$. The former is a stochastic process that has little variation. Whereas the latter is a deterministic process. Hence whilst the quantitative difference is small, the effect on the subsequent interpretation is huge. As a result, a thorough understanding of how to fit an SSM is essential to ensure that the model is informative, but yet still tractable and statistically robust.

4.3.2 Fitting an SSM

The objective of fitting an SSM is to estimate either the states $\mathbf{x}_t$, the model parameters θ, or a combination of the two. In the context of cell migration, the model parameters are rarely known and are often the primary focus of the model. However, the model parameters are estimated by maximising the likelihood of the model with respect the data. Yet the states form part of the data and they are unknown. Hence the states must also be estimated along with the

model parameters. Therefore fitting an SSM involves maximising the joint likelihood of the model parameters and the states conditional on the observations, $L(\theta, \mathbf{x}_{1:T}|\mathbf{y}_{1:T})$ (Shumway and Stoffer, 2011).

Expectation maximisation algorithm

In practice maximising the joint likelihood, $L(\theta, \mathbf{x}_{1:T}|\mathbf{y}_{1:T})$, becomes intractable for non-trivial problems. Thus a two-stage iterative process is known as the expectation-maximisation (EM) algorithm is adopted (Roweis and Ghahramani, 1999). The E stage estimates the state values from the conditional distribution of the states given the current parameter values of the model and the observed data $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \theta)$. The M stage then maximises the marginal likelihood of the model conditional on the observed data, $L(\theta|\mathbf{y}_{1:T})$. Note that the observed data $\mathbf{y}_{1:T}$ is dependent on the states $\mathbf{x}_{1:T}$ as defined in Equations 4.1 and 4.2 (Figure 4.3.1). This process is then repeated and continues until the likelihood of the model converges onto a local maximum. In short, the E stage generates estimated values of the states. The M stage then uses the estimated states to estimate the model parameters. The new model parameters are then used to generate new estimates of the states, and so forth, until the local maxima has been found.

Whilst the EM algorithm alleviates the strain of maximising the joint likelihood, $L(\theta, \mathbf{x}_{1:T}|\mathbf{y}_{1:T})$. Evaluating the marginal likelihood still requires evaluating the states from the conditional distribution, $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \theta)$. This means maximising the likelihood function of states over a $T \times T$ space which involves computing a T dimensional integral. Hence to reduce the computational expense the marginal likelihood, $L(\theta|\mathbf{y}_{1:T})$, is normally either approximated or, in specific cases, it is calculated directly via the Kalman filter (Kalman, 1960).

Kalman filter

The Kalman filter is recursive algorithm that can be used to estimate the states and marginal likelihood of an SSM with Gaussian errors. The filter operates in an iterative manor to update the mean and variance of the state up to time $t-1$ before then estimating the state value at time t. If an observation exists at time t the filter will then use this information to update and

improve its state estimate at time t. The process then repeats for the next state estimate at time $t+1$ (Kalman, 1960).

The first step of the Kalman filter is to initialise the mean and variance of the prior state distribution at time t_0, x_0. In this thesis the prior state distribution x_0 is set with a mean of 0 and an independent variance of 5, $x_0 \sim N(0,5)$. Whilst the parameters of the prior state distribution could be included as fixed unknowns, this increases the number of parameters that then need to be estimated by the model. Hence the diffuse prior enables the model to retain a high degree of flexibility whilst still minimising the number of unknown parameters.

In a simple SSM with no weight vectors or covariates the subsequent state distributions at time $t = 1...T$ are estimated by the following where $\mathbf{x}_{t|t-1}$ and $\mathbf{P}_{t|t-1}$ are the predicted state estimate and predicted estimate covariance respectively at time t:

- *Transition step*

    1. Predicted state estimate: $\mathbf{x}_{t|t-1} = \mathbf{B}\mathbf{x}_{t-1|t-1}$
    2. Predicted estimate covariance: $\mathbf{P}_{t|t-1} = \mathbf{B}\mathbf{P}_{t-1|t-1}\mathbf{B}^T + \mathbf{Q}$

- *Update step*

    3. Observation residual: $\mathbf{r}_t = \mathbf{y}_t - \mathbf{Z}\mathbf{x}_{t|t-1}$
    4. Observation covariance: $\mathbf{S}_t = \mathbf{Z}\mathbf{P}_{t|t-1}\mathbf{Z}^T + \mathbf{R}$
    5. Optimal Kalman gain: $\mathbf{K}_t = \mathbf{P}_{t|t-1}\mathbf{Z}^T\mathbf{S}_t^{-1}$
    6. Update state estimate: $\mathbf{x}_{t|t} = \mathbf{x}_{k|k-1} + \mathbf{K}_t\mathbf{r}_t$
    7. Update estimate covariance: $\mathbf{P}_{t|t} = (\mathbf{I} - \mathbf{K}_t\mathbf{Z})\mathbf{P}_{t|t-1}$

The transition step predicts the state estimate, $\mathbf{x}_{t|t-1}$, and estimate covariance, $\mathbf{P}_{t|t-1}$, at time t based on the values up to and including time $t-1$. Then, if an observation is missing at time t the filter will move on to the next iteration and repeat the transition step for time $t+1$. The ability to naturally handle missing observations is a powerful feature of the Kalman filter that makes it extremely useful in settings where incomplete data is common e.g. biology.

However, if an observation is present at time t the Kalman filter will then update the state estimate and estimate covariance to account for the information gained by the observation.

The update step begins by first calculating the difference between the actual observation at time t, $\mathbf{y}_t$, and the estimated observation, $\mathbf{Z}\mathbf{x}_{t|t-1}$. The observation covariance is then calculated in step 4 akin to the estimate covariance in step 2. The filter then compares the estimate covariance against the observation covariance to determine the magnitude of the update needed to both the state estimate and estimate covariance, known as the Kalman gain (Kalman, 1960).

Kalman gain

The Kalman gain is pivotal to the Kalman filter because it draws together both sources of variation within the model: the state and the process. The magnitude of the Kalman gain then dictates the subsequent improvement to both the state estimate and estimate covariance in steps 6 and 7.

The behaviour of the Kalman gain can be understood more clearly by expressing it such that:

$$\mathbf{K}_t = \frac{\mathbf{P}_{t|t-1}\mathbf{Z}^T}{\mathbf{Z}\mathbf{P}_{t|t-1}\mathbf{Z}^T + \mathbf{R}} \tag{4.3}$$

Hence when the observation model variance, $\mathbf{R}$, is small the Kalman gain tends to $\frac{1}{\mathbf{Z}}$. Thus when substituted into stages 6 and 7 of the Kalman filter the updated state estimate and state covariance depend primarily on the observation. That is, when the observation is accurate, the filter relies on the observation to estimate the state.

However, when the predicted estimate covariance, $\mathbf{P}_{t|t-1}$, is small the Kalman gain tends to 0. As such the updated state estimate and estimate covariance depend primarily on the state of the model. That is, when the state does not vary much through time, the model ignores the observation and relies on the value from the previous state.

Kalman filter results

The temporal nature of the Kalman gain means that its magnitude, and therefore its effect, can change dynamically over the course of the observed time series. This allows the Kalman filter to handle unexpected disturbances within the data without it impacting on the estimation performance of the underlying states. In fact, the performance of the Kalman filter is so robust that the Kalman filter estimates of the states are also the maximum likelihood state estimates.

The final part of the E step is to re-run the Kalman filter in reverse and combine the results from both directions to ensure that the temporal order is not biasing the estimates of the states, known as the Kalman smoother. The smoothed states are then used in the M stage of the EM algorithm to estimate the fixed parameters in the model (Shumway and Stoffer, 2017).

4.3.3 Model selection

The EM algorithm can be used to estimate the parameters of an SSM by combining the Kalman filter with maximum likelihood estimation. However, an SSM can be defined in multiple different forms due to the flexibility in the parameter structure (Section 4.3.1). Thus, a process of model section is needed to determine which parameter combination best fits the observed data (Siple and Francis, 2016).

Population	Ancestor	Escape	Invasion	Colonisation	Total
Sample size	24	58	105	57	244

Table 4.1: The final data set in Chapter 4 stratified by population.
Displayed are the number of cells within each experimentally evolved population that collectively form the final data set within Chapter 4. The data set is a subset of the 813 cells that were modelled in Chapter 3. Yet, in addition to the previous selection criteria, a cell also needed to have present for at least 8 of the 12 hours of tracking and not been involved in a cell division event.

In this chapter the data set is a subset of the 813 cells that were modelled previously (Chapter 3). In addition to the previous selection criteria a cell also needed to satisfy the

following two conditions. Firstly, a cell must not have divided during the tracking period or be the daughter cell of a division event. This is to prevent the model from trying to estimate the underlying state of a cell that is known to no longer exist. Secondly a cell needed to have been successfully recorded for at least 8 of the 12 hours of tracking. Whilst this condition is more subjective, it ensures that the model is fitted to a data set in which more values are present than missing. A total of 244 cells met these conditions and together they form the final data set (Table 4.1).

Optimal number of factors

The first step in dynamic factor analysis is to determine the optimal number of factors to include in the model. Ideally this is performed by fitting n different covariate free models where the number of factors increases sequentially from $1...n$. The optimal number of factors are then chosen based on the model that has the smallest AICc value (Zuur et al., 2003). Unfortunately, the 244 cell data set in this chapter means that there are two issue with adopting this approach.

Firstly, in a time series context, the 244×360 cell migration data set is extremely large. A total of 244 separate models would need to be fitted for each variance covariance structure. Also, as the number of factors in the model increases, the corresponding time taken to fit each model increases dramatically. As a result, fitting a model with even a moderate number of factors, $m > 7$, quickly becomes intractable.

Secondly, the number of factors need to reflect the underlying biological structure of the system. Thus, fitting a 10 factor model to a data set containing 4 experimentally evolved populations of cells would lack clear biological justification. As a result, in this chapter, the optimal number of factors were chosen by comparing models with a factor total ranging from 1...4. This meant that in the extreme case a model could still be estimated with a unique factor for each experimentally evolved population. Similarly, an independent parameter was estimated for each state in the state matrix allowing the state to progress as a random walk or a mean reverting process.

Variance structure

To ensure that the model remains identifiable the state variance, **Q**, has to be set equal to the identity matrix, **I** (Zuur et al., 2003). However, the observation variance, **R**, is undefined and therefore needs to be selected. In this chapter the following 6 different structures were compared:

1. Identity matrix (no estimation)
2. A single global estimate (1 estimated parameter)
3. Independent estimates for each video (33 estimated parameters)
4. Independent estimates for each cell (244 estimated parameters)
5. Correlated estimates for each video (66 estimated parameters)
6. Correlated estimates for each cell within a given video (1265 estimated parameters)

Across the 24 different model combinations the AICc value was minimised by a 4 factor model with correlated estimates for each video (Table C.2, Appendix C). The need for correlated estimates in each video maybe due to the cell density within each video affecting the quantification of the cell morphology. In short, if the video has a high density of cells in frame t then the density is also expected to be high in frame $t+1$. In turn, the high density of cells may then mean that extracting the morphology of multiple cells is more error prone. As a result, the video has a high observation variance that is also temporally correlated.

Covariate selection

Next, the inclusion of known covariates such as the speed of migration, nearest neighbour distance, and then interaction of the two were tested to determine whether the model error could be further reduced. To ensure that the model remains identifiable the covariates are strictly limited to the observation model, **d**, (Zuur et al., 2003). Likewise, to remain tractable the covariates are assumed to be observed without error and without any missing values.

However, the speed of migration and nearest neighbour distance do contain missing values as a result of the tracking process (Section 3.2.2). Hence before they can be used as covariates the missing values need to be imputed. To ensure that the imputed values are as accurate as possible a dynamic factor model was fitted to each covariate data set. The model selection process was the same as the covariate free model. The speed of migration model had 4 factors and a correlated observation variance structure for each video, akin to the covariate free model (Table C.3, Appendix C). The nearest neighbour model had 4 factors and a correlated observation variance structure for each cell within a given video (Table C.4, Appendix C). The missing values in each data set were then imputed from the estimated state predictions generated by the Kalman smoother.

The imputed covariates were then tested through a process of forward selection to determine whether they improve the model fit by lowering the AICc. The number of factors were set equal to 4 and each covariate addition was evaluated across the 6 different observation variance structures. In the first instance, the covariate effects were estimated at a population level (Table C.5, Appendix C) before the same process was then repeated at a single cell level (Table C.6, Appendix C). In both cases all 3 covariates were found to reduce the model AICc with independent estimates in the observation variance structure. Hence a full model was estimated at both the population and single cell level. Finally, prior to post-hoc analysis, the model was re-run at 10 different starting conditions to ensure that the global minima was found.

4.4 Results

4.4.1 Short term phenotypic flexibility

Time invariant population model

To act as a time invariant reference a linear mixed model was first fitted to the 244 cell data set. The model was fitted in accordance with the same model selection process detailed in Section 3.4.2 where covariates were chosen through a process of forward selection. The

populations were included as a fixed effect and then the intercepts and slopes were allowed to vary between each population. The significant parameters were then used to fit a reduced model to each population (Figure 4.3).

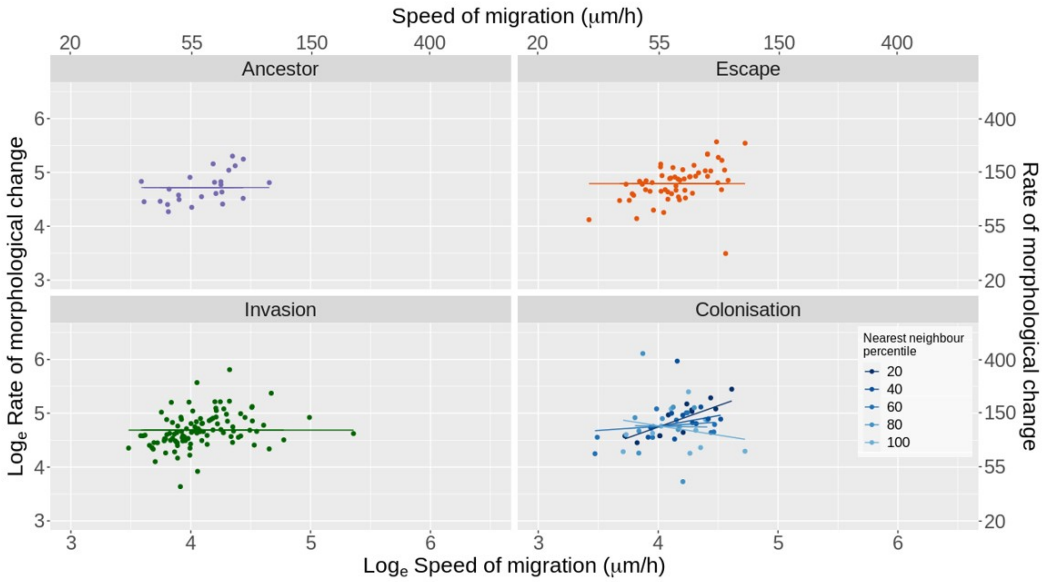


Figure 4.3: The rate of morphological change against the speed of migration. The natural log-transformed rate of morphological change plotted against the natural log-transformed speed of migration. The straight lines represent the reduced model for each treatment using only parameters that are significant at the 5% level. The ancestor (N = 24), escape (N = 58), and invasion (N = 105) populations have an intercept-only model fitted. The speed of migration ($p = 3.264 \times 10^{-3}$), the distance to the nearest neighbouring cell ($p = 1.572 \times 10^{-2}$) and the interaction of the two ($p = 1.133 \times 10^{-2}$) was significant in the colonisation population (N = 57). The shaded lines indicate the nearest neighbour percentile.

In the ancestor populations neither the speed of migration nor the distance to the neighbouring cells significantly affected the rate of morphological change. As such an intercept only model was fitted to the data. Also, in contrast to the previous chapter, neither the speed of migration nor the distance to the neighbour cells significantly affected the rate of morphological change in either the escape or invasion populations. Hence an intercept only model was also fitted to both populations. However, in the colonisation populations, the model remained the same as in Chapter 3. The rate of morphological change was dependent on the speed of migration, the distance to the nearest neighbouring cell, and the interaction

of the two: as the distance between neighbour cells increases, the relationship between the rate of morphological change and the speed of migration becomes negative. To ensure that the results were not affected by a small cluster of potential outliers the same analysis was repeated after having removed any influential data points (Figure B.3) defined by a Cook's distance > (4 / N) where N is the sample size (Bollen and Jackman, 1985). The same qualitative relationship was still present.

Population state space model

Next, the morphological behaviour was evaluated temporally with covariate effects estimated at a population level. This meant that 4 parameters were estimated for each covariate similar to a linear regression model. To test whether the model performed equally well for each population the root mean squared error (RMSE) was calculated for each cell based on the one-step-ahead-residuals (Harvey, 1990). An analysis of variance (ANOVA) was then used to compare the average RMSE between each experimentally evolved population. There was no significant difference in the average goodness of fit between the 4 experimentally evolved populations.

In the ancestor, invasion, and colonisation populations the speed of migration, nearest neighbour distance, and the interaction of the two were all significant at a 5% level (Figure 4.4). However, in the escape populations the speed of migration and interaction were significant but not the nearest neighbour main effect as seen by the 95% confidence interval overlapping 0 (Figure 4.4). Hence in the escape populations the nearest neighbour distance does not significantly affect the rate of morphological change when an individual cell is moving at its average speed of migration.

The change in significant population covariates between the linear regression model (Figure 4.3) and the population state space model (Figure 4.4) highlight that all of the cells can adopt similar phenotypic behaviours. Yet the frequency at which a given behaviour is adopted varies between different populations. In the linear regression model the covariates are calculated as the average value over the 12 hour migration period. Thus on average the ancestor, invasion, and escape populations do not significantly vary their rate of morpholog-

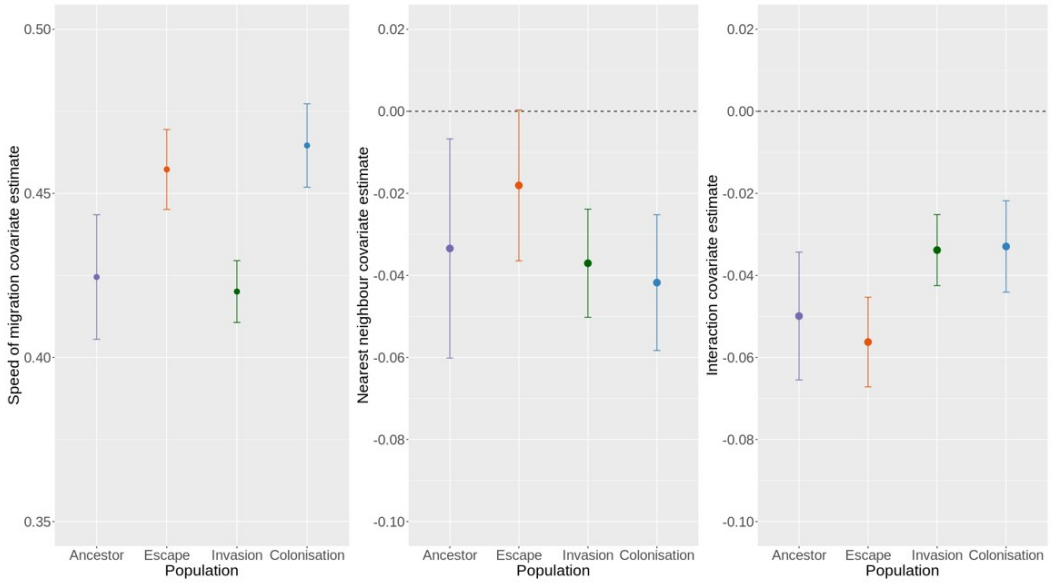


Figure 4.4: The state space covariate estimates at a population level and corresponding 95% confidence interval.
A plot of the covariate estimates for each experimental population within the population level state space model. The centre dot signifies the covariate estimate and the error bars are 95% confidence intervals. A covariate is significant within a given population if the 95% confidence interval does not overlap 0, as seen by the black dotted line. The ancestor, invasion, and colonisation populations can be seen to have a significant speed of migration, nearest neighbour and interaction effect. In contrast, the escape populations have a significant speed of migration and interaction, but the nearest neighbour main effect is not significant as seen by the 95% confidence interval overlapping 0 (95% CI = $[3.257 \times 10^{-4}, -3.645 \times 10^{-2}]$).

ical change in response to their average speed of migration or nearest neighbour distance. However, importantly, this does not mean that their rate of morphological change never varies in response to their speed of migration or nearest neighbour distance. The significant covariates in the population state space model highlight that the rate of morphological change can vary in response to the speed of migration and nearest neighbour distance. In contrast, the colonisation populations display the same phenotypic behaviour in both the linear regression model and population state space model. Hence this suggests that the colonisation populations vary their rate of morphological change more often in response to the speed of

migration and nearest neighbour distance compared with the other experimentally evolved populations.

4.4.2 Dispersal heterogeneity

Single cell state space model

Finally, the morphological behaviour was evaluated temporally with covariate effects estimated on a cell specific basis. In contrast to the 2 previous models this meant that 244 parameters were estimated for each covariate. To ensure that the model performed equally well across each of the 4 populations the RMSE was calculated for each cell based on the one-step-ahead-residuals (Harvey, 1990). An analysis of variance (ANOVA) was then used to compare the average RMSE between each experimentally evolved population. There was no significant difference in the average goodness of fit between the 4 experimentally evolved populations.

In the cell specific state space model, a variety of different phenotypic behaviours were detected across each of the 4 experimentally evolved populations. The spectrum of behaviours ranged from cells in which none of covariates were significant at a 5% level through to cells in which all of the covariates were significant (Figure 4.5). Yet, 98% of the cells where characterised by one of the following 4 covariate combinations and thus they remain the focus for further analysis:

- Speed of migration only
- Speed of migration and nearest neighbour distance
- Speed of migration and the interaction
- Speed of migration, nearest neighbour distance, and the interaction.

The most frequent covariate combination across all 4 of the experimentally evolved populations was the speed of migration only, accounting for 54.4% of cells. The speed of migration and nearest neighbour combination accounted for 15.3% of the 244 cells and the speed of migration and interaction accounted for 23%. Finally, the speed of migration, nearest neighbour, and interaction (known as the full model) accounted for 4.9% of the 244 cells.

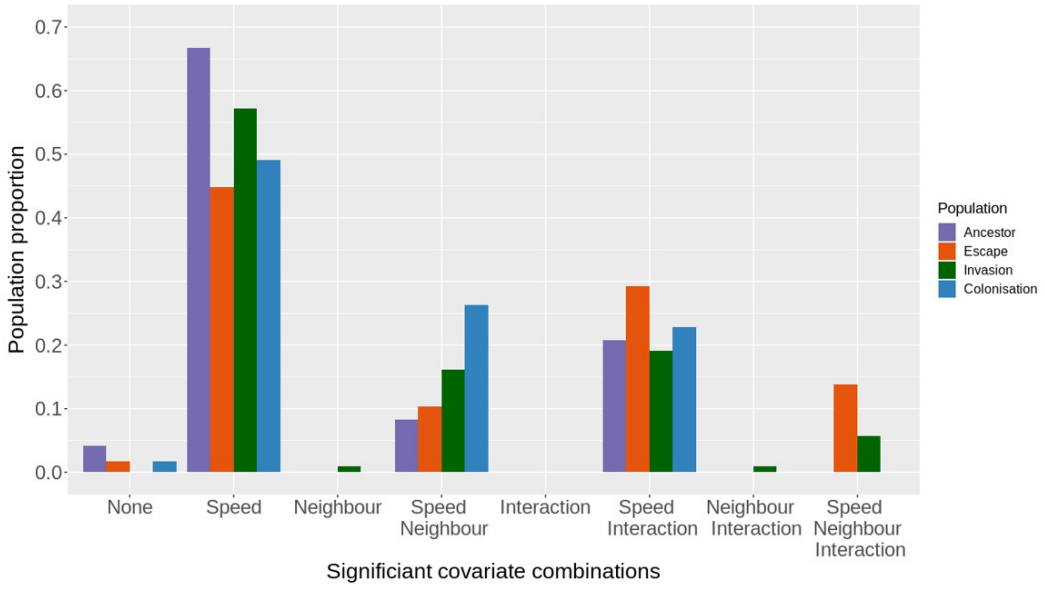


Figure 4.5: The proportion of cells within each experimental population that have a significant covariate combination within the single cell state space model.
A plot of the proportion of cells within each population that have a significant covariate combination when the covariates are estimated at a single cell level. A covariate is significant if the 95% confidence interval for that cell does not overlap 0. The cells are then stratified according to the significant covariates and population type. As a result, each strata is independent such that an ancestor cell with a significant speed of migration and nearest neighbour covariate effect cannot also be counted in the speed of migration only strata.

Population specific covariate combination

A Chi-Squared test was used to test whether there was a significant association between the experimental population and the specific covariate combination. There was a significant association between the experimental populations and the different covariate combinations at a 5% level ($p = 0.01393$, $N = 239$). A post-hoc Bonferroni multiple comparison test was then used to identify which populations were significantly associated with a given covariate combination. The escape populations were found to have a significantly larger proportion of the population associated with a full model of covariate effects ($p = 0.04143$, $N = 239$). The full model of covariates accounted for 13.8% of the cells in escape populations compared to 5.7% of the cells in the invasion populations. However, in contrast to population level models, none of the cells in the colonisation populations had a full model of significant covariates.

Nevertheless, the colonisation populations did have the highest proportion of cells, 26.3%, in which the speed of migration and nearest neighbour main effects were significant but not the interaction, a phenotypic behaviour not seen at the population level.

4.5 Discussion

The individual time series recorded for each experimentally evolved cell during migration was utilised in this chapter to investigate whether high levels of phenotypic flexibility are associated with selective pressures akin to later stages of metastasis. Then, once the degree of phenotypic flexibility had been partition, the level of within population phenotypic heterogeneity was evaluated to determine whether levels of heterogeneity were elevated in cells selected for distant site colonisation.

4.5.1 Elevated sensitivity in a crowded environment

The spatial and temporal heterogeneity at a distant site expected to be is less than within the primary tumour or the local tumour micro-environment. The combination of an established vasculature network and a tightly controlled population size means that excess resources are expected to be scarce (Pries et al., 2001). A similar selective pressure can be seen within the colonisation assays and it may explain why cells selected for distant site colonisation vary their morphological behaviour more frequently in response to environment changes.

In the ancestor, escape, and invasion populations the cellular cohort is periodically partitioned and then a subset of cells are moved into a new environment. In contrast, the colonisation populations remain as an entire cohort. As a result, if resource levels are updated evenly, then the colonisation populations are expected to spend a longer duration in a state of resource poverty owning to the larger average population size. Hence the scarce supply of resources may then select for cells that can maintain a high level of vigilance to ensure that any available resources are detected, and then captured.

Similarly, the non-significant neighbour main effect in the escape populations may also be a result of selection driven by changes in population density. Initially the local cell density

is high within the escape assays due to the cells being tightly packed into the high density collagen core. Then, over time, the cells escape outwards into the low density collagen surroundings. This causes the local cell density to reduce before eventually increasing due to the cells actively proliferating. A similar dynamic can be seen within the invasion assay, albeit the cells are invading inwards towards the Matrigel core. However, importantly, the rate of local cell density is expected to decrease faster within the invasion assay compared to the escape. This is due, in part, to the smaller average distance from the selection boundary, the margin at which the two environments join, and that the cells within the invasion assay are able to migrate outwards in the opposite direction. As a result, the slower reduction in population density within the escape assay means that on average a cell will have to wait longer until its resource availability increases. This may therefore mean that selection acts on a cell to increase its long term survival and only respond to environmental cues when absolutely necessary i.e when it is either trapped or migrating at high speeds.

4.5.2 Spatial heterogeneity selects for multiple dispersal strategies

The prolonged period of resource scarcity within the colonisation populations may mean that the spatial resource heterogeneity is larger compared to the corresponding temporal resource heterogeneity. Hence, in such an environment, ecological dispersal theory would predict that multiple conditional dispersal strategies can exist at an evolutionary optima (McPeek and Holt, 1992). Thus, this may explain why none of the cells have a significant full model of covariate effects but why a quarter of the cells have a significant speed significant speed of migration and nearest neighbour effect without a significant interaction.

The significant nearest neighbour main effect means that the sub-population of cells can continuously detect subtle changes in environmental conditions. Hence when the local cell density is high, the individual cell may benefit from a constant level of resource surveillance. In contrast, if the local cell density is low, then the cost of constant detection might outweigh the corresponding benefit causing a reduction in fitness. Hence this would explain why the behaviour is less frequent within the other experimental populations where the cell density is periodically reduced. Yet in the colonisation populations the combination of high cell density

and a temporally constant environment means that the behaviour can be selected. If true, this could have important implication because it would suggest that the diversity within the colonisation population is higher and therefore the corresponding rate of evolution is also likely to be higher.

4.5.3 Frequency of behaviour is as important as the behaviour itself

In summary, modelling temporal changes in cancer cell morphology provides fresh insight into the frequency at which phenotypic behaviours are displayed within breast cancer cells that are experimentally selected for different stages of the metastatic cascade. Likewise modelling phenotypic behaviour at a single cell level highlights the pervasive heterogeneity that arises from different ecological selective pressures. A future next step will be to determine whether the covariate interactions in each experimental population occur at the same relative level of migration speed or nearest neighbour distance. That is, do cells selected for their ability to escape switch their phenotypic behaviour at a lower relative speed compared with cells selected for invasion due to the different rates of reduction in population density. Furthermore, applying sequential selective pressures will be important to determine whether the speed of migration and nearest neighbour covariate combination is retained within cells selected for distant site colonisation. If so, then it may highlight a key phenotypic behaviour that is essential for metastatic success.

Chapter 5

Poly-aneuploid cancer cell specific signalling

5.1 Introduction

The majority of cancer cells are aneuploid (Taylor et al., 2018), they possess an unbalanced chromosomal complement (Ben-David and Amon, 2020). Mistakes during cell division cause changes to occur in both the chromosomal structure and number (Gordon et al., 2012). This leads to increased genetic diversity within the population akin to the effect of a mutation. However, due to the number of genes that are affected, chromosomal aberrations typically create a much larger genetic disturbance. As a result, the corresponding impact on the fitness of a cell is often more dramatic leading to increased rates of phenotypic adaption (Sansregret and Swanton, 2017).

Polyploidization

An extreme form of structural disturbance associated with worse patient outcomes is seen by the duplication of the entire chromosomal complement, known as polyploidization (López et al., 2020). Polyploidization is seen in multiple different cancer types and is commonly assumed to be an early evolutionary event in cancer progression to overcome the accumulation of deleterious mutations (Bielski et al., 2018). However, on a cellular

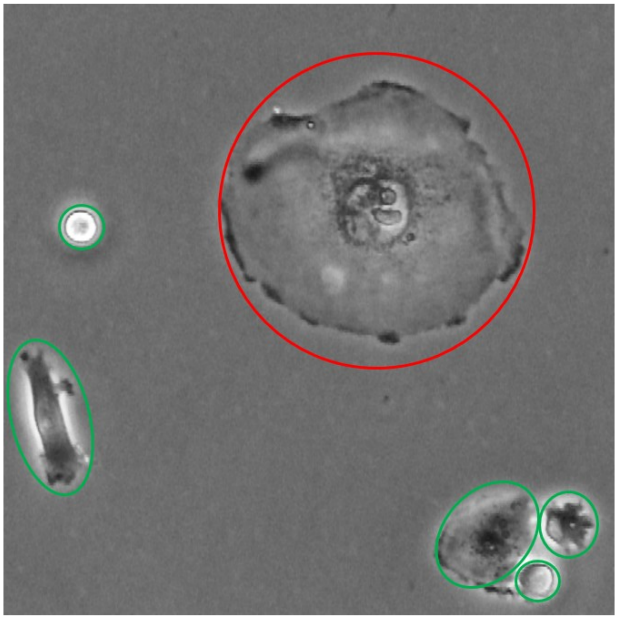


Figure 5.1: A phase contrast image of a poly-aneuploid cancer cell (PACC).
A phase-contrast image of a poly-aneuploid cancer cell (PACC) taken from the 12 hour time-lapse videos that were collected in Chapter 3. A large PACC can be seen within the centre of the image (circled in red). In contrast, 5 smaller normal cancer cells can be seen within the local vicinity (circled in green).

timescale, polyploidization also leads to the formation of a physically large, but highly motile, quiescent cell type known as a poly-aneuploid cancer cell (PACC) (Pienta et al., 2020) (Figure 5.1). Importantly, PACCs have been found within the distant site metastases of patients that have subsequently died from metastatic cancer (Mannan et al., 2020). Hence this would suggest that duplication events maybe a recurrent feature in cancer development and that the formation of PACCs maybe associated with a lethal stage of cancer progression.

5.1.1 Polyploid formation and function

Whilst PACCs are cancer specific, polyploidization also occurs within a range of different somatic tissues (Fox et al., 2020) through two independent processes: cellular fusion and deviations in DNA duplication (Aguilar et al., 2013; Calvi, 2013). An osteoclast is an example of a cell fusion derived somatic polyploid cell that forms from the fusion of bone marrow macrophages. In turn, osteoclasts are then responsible for the degradation of bone to

release essential minerals such as calcium (Pienta et al., 2020). In contrast, a megakaryocyte is an example of a somatic polyploid cell that forms due to deviations in DNA duplication. Specifically, a megakaryocyte is formed through a process known as endomitosis whereby the chromosomes in a cell divide, but the nucleus remains as a single organelle (Orr-Weaver, 2015). Once formed, a megakaryocyte is then responsible for the production of platelets, a key component of blood that is necessary for clotting (Zimmet and Ravid, 2000). The wide range of physiological functions that are performed by polyploid cells demonstrate the importance of ploidy increases and the need to evaluate the exact function of PACCs in cancer progression.

Therapeutic resistance

Experimental studies have found that duplication events, leading to the formation of PACCs, can be induced in response to adverse environmental conditions such as hypoxia (Lopez-Sánchez et al., 2014) and exposure to cytotoxic drugs (Lin et al., 2019). In turn, once the environment conditions improve, a proportion of the PACC population are seen to de-polyploidize back into viable proliferative cancer cells (Erenpreisa et al., 2011). However, after the de-polyploidization event, the PACC progeny display an increased tolerance to the environment pressures that induced the initial PACC formation (Puig et al., 2008). That is, if a PACC forms in response to a cytotoxic agent, then the PACC progeny display a higher cytotoxic tolerance compared to the pre-PACC cancer cells. As a result, this would suggest that chromosomal duplication maybe an active strategy that some cells can adopt to increase their chances of survival within a hostile environment (Mallin et al., 2020). Hence finding, and then targeting, PACC vulnerabilities could prove to be an essential step in combating metastatic progression and therapeutic resistance.

Migration difficulties

The increase in cell size from polyploidization is known to cause a corresponding decrease in the surface area to volume ratio of a polyploid cell (Marshall et al., 2012). In turn, this maybe a critical feature in the PACC defence strategy by lowering the overall toxin load that a PACC

will experience compared to a normal cancer cell (Pienta et al., 2021). Yet the increase in size also means that the metabolic rate of a PACC is higher compared to a normal cancer cell (Coward and Harding, 2014). Thus, the level of resources needed for PACC survival are also expected to be higher compared to a normal cancer cell. However, in the context of resource detection, the reduction in surface area to volume ratio may also prove to be problematic during PACC dispersal.

Firstly, morphological changes associated with resource detection carry an energetic cost that is expected to scale with cell size. Thus, the process of resource detection is likely to be more expensive in a PACC compared to a normal cancer cell. Furthermore, the reduction in surface area to volume ratio means that a PACC is expected to change its morphology more often to achieve the same relative level of resource detection per unit volume. In short, resource detection is expected to be a more costly process for a PACC, but a PACC also needs to participate in resource detection more often. Secondly, if a resource deposit is detected, a PACC's migration through the complex tumour microenvironment maybe restricted by its enlarged nucleus (Wolf et al., 2013). Thus, a PACC maybe forced to participate in extensive environmental rearrangements incurring a further energetic penalty.

The conductors of group cooperation

Alternatively, the prolonged survival time of a PACC may mean that it could benefit from co-operating with normal cancer cells within the local vicinity. For example, somatic polyploid cells are known to down-regulate metabolic pathways associated with oxidative phosphorylation (Vazquez-Martin et al., 2016). As a result, polyploid cells are forced to participate in aerobic glycolysis which in turn releases diffusible metabolites into the local microenvironment as a by product (Nakajima and Van Houten, 2013). Aerobic glycolysis is also known to be a primary form of energy production in cancer cells, especially PACCs (Donovan et al., 2014). However, cancer cells are known to induce aerobic glycolysis in neighbouring stromal cells as a way to obtain vital metabolites and then benefit from more energy efficient oxidative pathways (Pavlides et al., 2009). Thus, if PACCs are producing a similar diffusible product then nearby cancer cells could be attracted to the high concentration of diffusible by

products subsequently benefit from the use of oxidative pathways. In turn, a PACC may then co-opt the highly energised cancer cells to initiate a leader-follower dynamic akin to the onset of cellular streaming during tumour invasion (Zhang et al., 2019). This would dramatically reduce the energetic cost to the individual PACC whilst also benefiting the neighbouring cancer cells by providing vital resources.

Chapter overview

To understand the role of PACCs in metastatic dispersal the chapter begins by characterising the phenotypic behaviour of a PACC during migration. In turn, a state space model is then built to test whether there is a different behaviour response from a cell when it neighbour is a PACC compared to a normal cancer cell. That is, do normal cancer cells detect when their nearest neighbour is a PACC, and then significantly change their morphological behaviour as a result, known as the PACC model. Finally, the results of the PACC model are then discussed before a series of open questions are proposed that need to be addressed to understand the role of PACCs in cancer progression.

5.2 PACC characteristics

To investigate the behaviour of PACCs, a data set is needed in which multiple PACCs exist. To ensure that the same data set can be used throughout the remainder of this chapter the data set must also include multiple cancer cells that have interacted with both a PACC and a normal cancer cell during their migration for the PACC model. As a result, the final data set in this chapter was curated through a 3 step process that started with the original 33 time-lapse videos that were collected in Chapter 3.

5.2.1 Data curation

The data curation process began by first identifying which of the 33 time-lapse videos contained a PACC. A cell was classified as a PACC if it had both an enlarged cellular and nuclear area, hence this includes both mono and multi nuclear PACCs. Two independent

Population	Ancestor	Escape	Invasion	Colonisation	Total
Sample size	24	0	35	10	69

Table 5.1: The final data set in Chapter 5 stratified by population.
Displayed are the number of cells within each experimentally evolved population that collectively form the final data set within Chapter 5. That data set is a subset of the 244 cells that were modelled in Chapter 4. In addition to the previous selection criteria, each time-lapse video also needed to contain a PACC and at least one cell in each video needed to have a PACC nearest neighbour during their migration. Further details regarding the data selection process can be found in Section 5.2.1.

operators then manually inspected each video and if the same cell was highlighted as a PACC by both operators then the cell was judged to be a PACC and thus the video was eligible. A total of 10 PACCs were identified across 8 of the 33 time-lapse videos. Invariably the manual selection process introduced a certain degree of subjectivity. However, due to the extreme size of a PACC relative to a normal cancer cell (Figure 5.1) the likelihood of a misclassification is expected to be low.

Next, the 8 time-lapse videos were then used to filter the 244 cell time series data set used in Chapter 4. This ensured that all of the cells within the subsequent PACC model were present within at least 8 of the 12 tracking hours and none of the cells were involved in a cell division event. A total of 76 cells were selected within the 8 time-lapse videos. Finally, across the 76 tracked cells at least one cell per video needed to have had a PACC nearest neighbour during their migration for the video to be included in the final data set. The final data set contained a total of 69 tracked cells in 7 time-lapse videos (Table 5.1). The 69 tracked cells included 5 PACCs and 25 cells in which their nearest neighbour had been a PACC during their migration.

5.2.2 Rate of morphological change scales with cell size

Once the data set had been curated, the rate of morphological change and the speed of migration were then compared between the PACC and normal cancer cell populations. However, owing to the small PACC sample size, 5 cells, it was not possible to quantify the migratory behaviour of the PACC sub-population directly. As a result, the rate of

morphological change and speed of migration were evaluated in response to the average cellular area. A linear model was then fitted to the normal cancer cell population, along with a 95% prediction interval, and then extrapolated to the PACC sub-population. In turn, if the 5 PACCs were found to be inside of the 95% prediction interval an inference could be drawn with respect to the cell area. Thus, enabling an implicit comparison to then be drawn due to the extreme difference in cell area between the PACC and normal cancer cell population. Alternatively, if the 5 PACC cells were found to be outside of the 95% prediction interval then this may suggest that a unique migratory behaviour is present within the PACC sub-population, possibly to compensate for the extreme size.

In the normal cancer cell population, the cell area was found to be significantly and positively correlated with the rate of morphological change at a 5% level ($p = 8.637 \times 10^{-12}$, β = 0.567, N = 64). The model was also found to explain a significant proportion of the variation within the average rate of morphological change (R^2 = 0.524). Hence within the normal cancer cells an increase in cellular area caused a corresponding increase in the rate of morphological change. Furthermore, the 5 PACC cells were also found to be inside of the extrapolated 95% prediction interval. Therefore, suggesting that the same positive correlation is also present within the PACC sub-population (Figure 5.2A). In contrast, the cell area was not significantly correlated with the speed of migration within the normal cancer cell population. Hence an intercept only model was fitted, which in turn included the 5 PACC cells within the extrapolated 95% prediction interval (Figure 5.2B). However, the intercept model explained only a small proportion of the variation (R^2 = 0) within the speed of migration. Therefore, suggesting that the speed of migration is either highly stochastic, or that it depends on factors not included in the model.

In summary, the results show that the rate of morphological change scales with cell area. Hence the rate of morphological change is expected to be considerably higher in the PACC sub-population compared to the normal cancer cell population owing to the extreme increase in cell area. In turn, to sustain the high rate of morphological change, a PACC is expected to have an increased metabolic rate meaning that the individual level of resources that a PACC needs to survive will be higher. However, the results also show that the speed

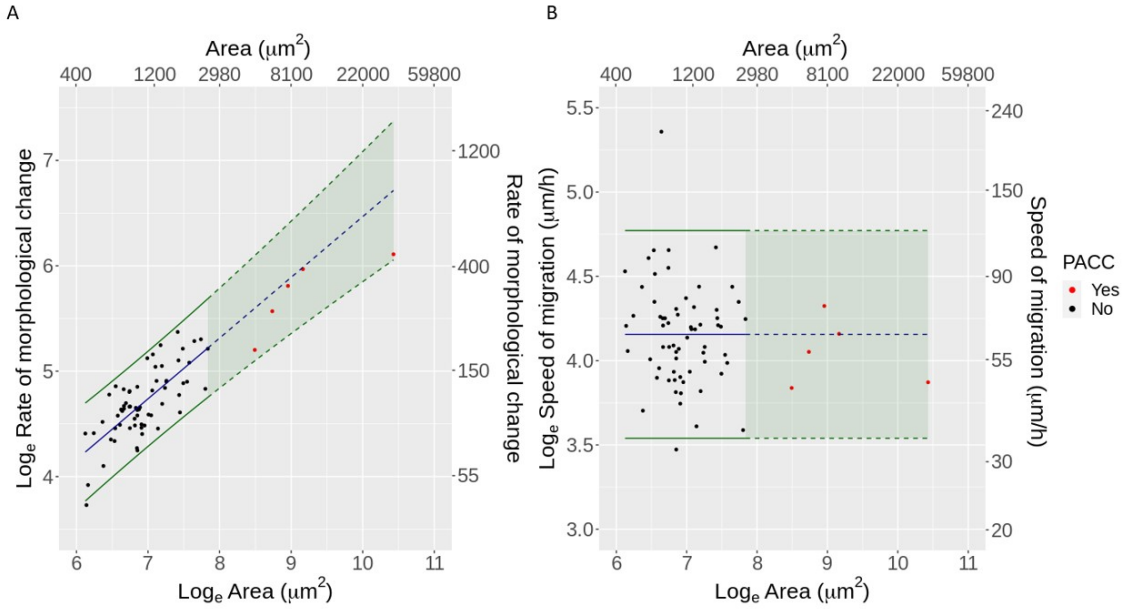


Figure 5.2: The effect of increased cellular area on migratory dynamics.
The solid blue lines represent the significant model in each data set and the solid green lines represent the corresponding 95% prediction interval. The dashed blue and green lines then represent the extrapolated model and 95% prediction interval respectively. Finally, the shaded green regions indicate the area enclosed by the extrapolated 95% prediction interval. (**A**) The natural log-transformed rate of morphological change plotted against the natural log-transformed cell area. The cell area was significantly and positively correlated with the rate of morphological change at a 5% level ($p = 8.637 \times 10^{-12}$, $\beta = 0.576$, N = 64). Hence an increase in cellular area caused a corresponding increase in the rate of morphological change. (**B**) In contrast, the cell area was not significant in the speed of migration and thus an intercept only model was fitted (N = 64). Finally, in both models, the 5 PACC cells were inside of the extrapolated 95% prediction interval indicating the same migratory dynamics were present within the PACC and normal cancer cell populations. As a result, this means that the PACC population is expected to have a higher rate of morphological change compared to the normal cancer cell population but a similar speed of migration.

of migration does not scale with cell area. Hence the speed of migration in the PACC sub-population is not expected to be faster than in normal cancer cell population. Whilst a PACC may need more resource to survive, it is not necessarily able to find the resources any faster than a normal cancer cell. Taken as a whole, this further supports the notion that a cooperative search strategy could be beneficial within the PACC sub-population. As a result, the following section discusses how a state space model can be used to evaluate whether

there is an interaction between a PACC and a normal cancer cell that might be indicative of a symbiotic relationship.

5.3 Cell - to - cell interaction model

Cell - to - cell interactions are dynamic and temporally dependent. Capturing transient changes in morphological behaviour requires the use of the individual time series data that is collected for each cell during tracking, as a discussed in Section 4.2. In turn, the time series data can then be modelled through the use of a state space model to account for the presence of autocorrelation (Durbin and Koopman, 2001). However, in contrast to the previous chapter, the purpose of the PACC neighbour model is to investigate acute temporal changes in morphological behaviour. Hence, the model has an increased focus on the underlying state of the system (Equation 4.1), and the dynamics that govern how an individual cell changes its behaviour through time. As a result, there are subtle but important changes to the model structure that need to be addressed prior to further analysis being conducted.

5.3.1 Model structure

To recap, a state space model partitions the variation within a time series in two separate models: state and observation. The state model is designed to reflect the underling dynamics of the system. The system is assumed to have a temporal structure that develops through time as a stochastic process. The latent state is then linked to the observed time series through the observation model that also accounts for the variability introduced by the sampling procedure (Section 4.3) (Durbin and Koopman, 2001).

In the previous chapter (Chapter 4) a state space model was used to quantify the short term phenotypic variability and heterogeneity in the signal processing behaviour of individual cells. The model was then used to evaluate whether increased phenotypic flexibility and heterogeneity was associated with selective pressures similar to late stages of the metastatic cascade. The model therefore focused on the variation between individual cells rather than the variation within a given cell. That is, the variation through time was acknowledged, but

was not the primary focus of the model. As a result, the temporal variation within each cell was approximated by a linear combination of 4 underlying generative processes, a technique known as dynamic factor analysis (Harvey, 1990; Zuur et al., 2003).

In contrast, the temporal variation within a given cell is the primary focus of the PACC model. The objective of the model is to detect whether there is a different behavioural response from a cell when its neighbour is a PACC compared with a normal cancer cell. That is, does the presence of a PACC transiently change the morphological behaviour of a neighbouring cell. As a result, the individual state of each cell needs to be modelled directly and thus the number of states in the model must equal the number of observations, $m = n$.

State matrix

The one - to - one relationship between the states and observations in the PACC model means that the observation matrix, $\mathbf{Z}$, must be set equal to the identity matrix, $\mathbf{I}$. In contrast, the structure of the state matrix, $\mathbf{B}$, is more flexible and needs be evaluated during the model selection process.

The state matrix, $\mathbf{B}$, is an $m \times m$ matrix that relates the current value of the state at time t, $\mathbf{x}_t$, to the previous value of the state at time $t-1$, $\mathbf{x}_{t-1}$. The structure of the state matrix is important because it characterises the underlying dynamics of the model and specifies how the model will develop through time. For example, if the state matrix is set equal to the identity matrix then the model will develop as a random walk. This means that the variance of the state distribution will increase through time, and hence the rate of morphological change for an individual cell may increase continually. Contrary to this, cellular behaviour is assumed to have an underlying level of persistence. Thus over an extended period of time the rate of morphological change for an individual cell is expected to revert back to an average value. Hence the state matrix of the PACC model is estimated such that the magnitude of each diagonal element is strictly less than 1.

Whilst the individual rate of morphological change for a given cell is expected to revert back to an average value, the time taken to revert back to the cell specific average is not necessarily constant across the population. For example, a change in morphology requires

the rearrangement of internal actin filaments (Lauffenburger and Horwitz, 1996; Olson and Sahai, 2008). A larger cell, such as a PACC, is therefore expected to have a greater number of filaments. Thus, in order to achieve the same degree of morphological change relative to a normal cancer cell a PACC needs to rearrange more filaments. In turn, the level of exertion needed to achieve the same relative change is expected to be higher and therefore deviations from the average rate maybe temporally shorter than in a normal cancer cell. The structure of the state matrix must therefore be carefully evaluated to ensure that the model accurately reflects the underlying dynamics of the system, akin to the observation variance structure in Chapter 4.

Covariate effects

A cell specific state also means that the covariate affects can be estimated on single cell basis in either the state or observation model. However, the effect, and subsequent interpretation of a covariate in either the state or observation model can be vastly different.

The observation model links the observed data to the underlying system dynamics whilst also accounting for the variability introduced by the sampling procedure. Yet, in certain systems, a significant proportion of the sampling variability maybe explained by the inclusion of known external factors. For example, if a change in whale population sizes is compared between different geographic regions then the water visibility within each region may affect the sample variability. That is, the sampling variation within a region with poor visibility is expected to be higher because it is harder to see the whales within the water. Hence the degree of water visibility maybe included as a covariate in the observation model to reduce the observation variance and therefore improve the model performance. In contrast, a covariate within the state model explains a significant proportion of the variation within the underling system dynamics. Thus in the whale example the number of fish within a given region maybe included as a state covariate because the fish stock will affect the underlying system dynamics irrespective of the water visibility. The focus of the PACC model is to characterise the underlying system dynamics and thus the covariate affects are strictly limited to the state model.

5.3.2 Model selection

A process of model selection was initially used to build a reference model that did not discriminate between neighbour types, known as the blind model. The first round of model selection sought to evaluate the structure of the state matrix, **B**, and state variance, **Q**. A further round of model selection was then used to evaluate the inclusion of common covariates such as the speed of migration and nearest neighbour distance. Note that the phrase *common covariates* is used throughout this chapter to signify covariates that are estimated for all 69 cells compared with covariates that are only estimated in a subset of cells.

Fixed observation variance

Theoretically the observation variance, **R**, could also be estimated. However, the small sample size (Table 5.1) meant that the model became unstable when estimating the observation variance and thus the observation variance had to be fixed. The magnitude of the observation variance for each cell was set equal to the average time-lapse video specific observation variance in Chapter 4. Whilst this could potentially bias the output of the model the only alternative, without collecting more data, was to assume that there was no observation variance and then remove the observation model. Taken as a whole, the fixed model variance appeared to better reflect the underlying knowledge of the system and thus a fixed variance was adopted for the remainder of this chapter.

Covariate free model

In the first stage of model selection the fixed observation variance was used to compare 3 different state matrix structures and 7 different state variance structures (Table 5.2). Across the 21 different model combinations the model AICc was minimised by an independent estimate for each cell in both the state matrix and the state variance (Table C.8, Appendix C). The need for a cell specific estimate in the state matrix suggests that there is significant variation between the individual cells in how their morphological behaviour develops through

time. Likewise, the cell specific estimate in the state variance also highlights that there is significant variation in the underlying system dynamics between individual cells.

Structure	Number of estimated parameters	State matrix	State variance
Identity matrix	0	⋆	✓
A single global estimate	1	✓	✓
Independent estimates for each population	3	✓	✓
Independent estimates for each video	8	⋆	✓
Independent estimates for each cell	69	✓	✓
Correlated estimates for each video	16	⋆	✓
Correlated cell estimates for each cell within a given video	432	⋆	✓

Table 5.2: The state matrix and state variance structures compared during model selection. The 3 different state matrix structures and the 7 different state variance structures compared during the covariate free model selection. A ✓ signifies that the structure was used for the given parameter where as a ⋆ signifies that it was not. Across the 21 different model combinations the model AICc was minimised by an independent estimate for each cell in both the state matrix and the state variance.

Full model

The next stage of model selection then evaluated whether the model error could be further reduced through the addition of known common covariates such as the speed of migration, nearest neighbour distance, and the interaction of the two. The covariates were assumed to be observed without error and without missing values akin to the previous chapter. As a result, the imputed covariate data in Chapter 4 was reused for both the speed of migration and the nearest neighbour distance for the remainder of this chapter.

The covariates were then tested at a single cell level through a process of forward selection to determine whether they improved the model fit by lowering the AICc. The addition of each covariate was evaluated with a fixed observation variance and a state matrix that had independent estimates for each cell. Importantly, each covariate was evaluated across the 7 different state variance structures to determine whether the optimal model state variance had

changed (Table C.8, Appendix C). All 3 covariates were found to reduce the model AICc and thus the blind model was fitted with a full set of common covariates at a single cell level with independent estimates in the state variance (Table C.9, Appendix C).

5.3.3 Neighbour identity

Once the blind model had been fitted, the PACC model was then fitted to evaluate whether normal cancer cells significantly change their morphological behaviour when their nearest neighbour is a PACC compared with a normal cancer cell. The PACC model is an extension of the blind model that also includes a neighbour identity covariate to highlight when the nearest neighbour cell at time t is a PACC.

The neighbour identity was generated for each of the 69 cells as a binary variable that equalled 1 when the nearest neighbour was a PACC and 0 when it was a normal cancer cell. Similar to the speed of migration and nearest neighbour distance the neighbour identity also contained missing values. However, in contrast to the common covariates, the neighbour identity was not imputed with a state space model. Instead, the imputed nearest neighbour distances in Section 4.3.3 were used to determine whether the missing neighbour identity at time t was a PACC. This approach was advantageous because it prevented the use of a non-Gaussian state space model and also ensured the neighbour identity corresponded with the nearest neighbour distance. Finally, a neighbour identity effect was estimated for the 25 cells that had a PACC nearest neighbour during their migration and the neighbour identity effect was set equal to 0 for the remaining 44 cells.

5.4 Results

5.4.1 PACC identity is significant

The neighbour identity covariate was found to improve the model performance and lower the PACC model AICc by 44 points compared to the blind model (Table 5.3). This suggests that

there is a significant change in morphological behaviour when the nearest neighbour cell is a PACC compared with a normal cancer cell.

Model	Number of estimated parameters	AICc	ΔAICc
Blind	345	58073	0
PACC	370	58029	-44
Alternative	362	58083	+10

Table 5.3: Neighbour identity model performance comparison.
Displayed are the performance for each of 3 models compared in Chapter 5. The blind model is a reference model that does not discriminate between neighbour types. The covariates in the blind model include the speed of migration, nearest neighbour distance and the interaction of the two. The PACC model is an extension of the blind model that also includes an identity covariate for the 25 cells which have a PACC nearest neighbour during their migration. Finally, the alternative model has the same structure as the PACC model but a cell has been chosen at random to be a pseudo PACC. The identity covariate in the PACC model was found to improve the model performance and reduces the model AICc by 44 points compared to the blind model. In contrast, the alternative model performed worse than the blind model and increased the model AICc by 10 points compared to the blind model.

Specific neighbours matter

To test whether the improved PACC model performance was an artefact of the increased model resolution an alternative model was also fitted. The alternative model randomly selected a cell from each of the 7 time-lapse videos to be a *pseudo* PACC. The same data curation process was then adopted as outlined previously (Section 5.2.1). The 69 cells within the alternative model contained a total of 3 PACCs and 17 cells in which their nearest neighbour had been a PACC during their migration. Whilst the number of cells that had a PACC nearest neighbour decreased by 68%, a Kruskal Wallis test found that there was no significant difference in the average time spent with a PACC nearest neighbour across the two models. The model was then fitted with an estimated neighbour identity effect for the 17 cells that had a PACC nearest neighbour and the neighbour identity effect was set equal to 0 for the remaining 52 cells.

In contrast to the PACC model, the neighbour identity covariate in the alternative model was found to reduce the model performance and increase the alternative model AICc by 10 points compared to the blind model (Table 5.3). This therefore suggests that the significant neighbour identity effect in the PACC model is capturing a unique biological phenomena. As a result, the output of the PACC model was then investigated to determine whether the temporal dynamics in the rate of morphological change were different among the 3 experimentally evolved populations.

5.4.2 Temporal phenotypic dynamics

To test whether the PACC model performed equally well for each of the populations the root mean squared error (RMSE) was calculated for each cell based on the one-step-ahead-residuals (Harvey, 1990). An analysis of variance (ANOVA) was then used to compare the average RMSE between each experimentally evolved population. There was no significant difference in the average goodness of fit between the 3 experimentally evolved populations.

Phenotypic behaviour is not population specific

The significant common covariate effects were then evaluated across the 69 cells to determine whether a specific morphological behaviour was associated with a given experimental population. A spectrum of different morphological behaviours were detected ranging from cells in which none of the common covariates were significant at a 5% level through to cells in which all of the common covariates were significant.

Across the 69 cells the most frequent common covariate combination was the speed of migration only accounting for 31.9% of the cells. The speed of migration and nearest neighbour combination accounted for 29% of the 69 cells and the speed of migration and interaction accounted for 20.3% of the cells. Incidentally, this is the same significant common covariate combination rank order as the previous chapter (Section 4.4.2). Finally, the speed of migration, nearest neighbour, and interaction accounted for 8.7% of the 69 cells and the remaining 10.1% of cells had no significant common covariate effects (Figure 5.3).

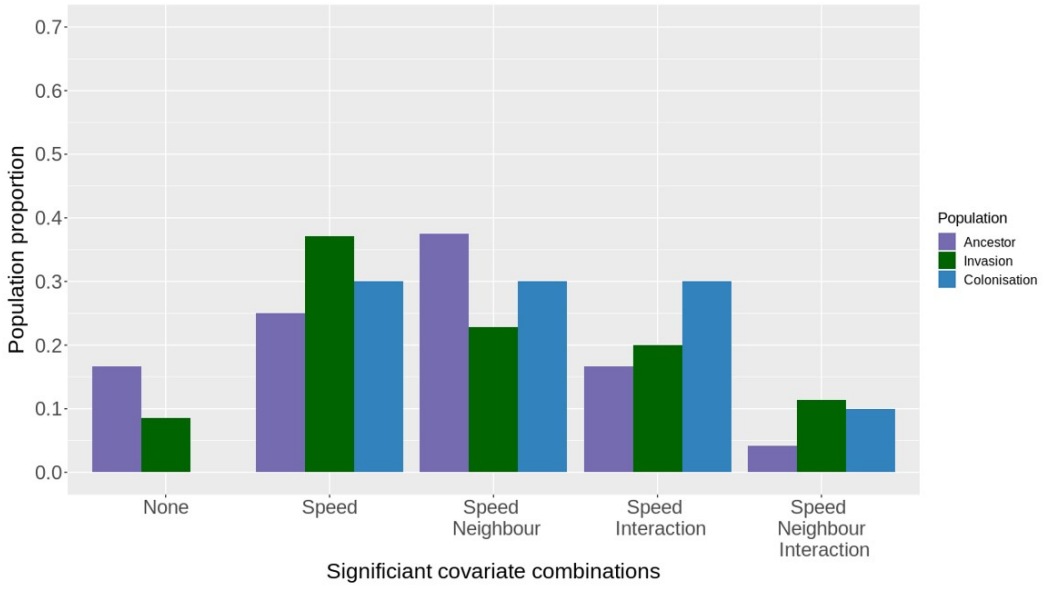


Figure 5.3: The proportion of cells within each population that have a significant common covariate combination with the PACC model. A plot of the proportion of cells within each population that have significant common covariate when estimated at a single cell level within the PACC model. A covariate is significant if the 95% confidence interval for that cell does not overlap 0. The cells are then stratified according to the significant covariates and population type. As a result, each strata are independent such that an ancestor cell with a significant speed of migration and nearest neighbour covariate effect cannot also be counted in the speed of migration only strata.

A Chi-Squared test was used to evaluate whether there was a significant association between the experimental population and the significant common covariate combination. There was no significant association between the experimental population the different common covariate combinations at 5% level. This means that within PACC model the significant common covariate combination is independent of the experimental population.

A significant PACC response is cell specific

Next, the significant neighbour identity effect was investigated. Across the 25 cells in which a neighbour identity effect was estimated 68% had a significant nearest neighbour response. That is, 17 of the 25 cells significantly changed their rate of morphological change in response to the distance from their nearest neighbouring cell. Furthermore, 47% of the 17 cells also

had a significant neighbour identity effect. That is, 9 of the 25 cells displayed a significant change in their rate of morphological change when their nearest neighbour was a PACC compared with a normal cancer cell.

The duration over which each of the 25 cells had a PACC nearest neighbour ranged from 0.28% to 99.7% of their migration trajectory. Likewise, the cell specific average distance from a PACC ranged from 1.01μm to 29.15μm across the 25 cells. The PACC nearest neighbour duration and distance were then tested to determine whether a significant neighbour identity effect was more likely in cells that spent a longer duration next to a PACC or were physically closer to a PACC. A Kruskal Wallis test was used to compare the mean PACC nearest neighbour duration and distance between cells with a significant or non significant neighbour identity effect. There was no significant difference in the PACC nearest neighbour duration or distance at a 5% level between cells in which a neighbour identity effect was significant or not. This therefore suggests that the neighbour identity effect is cell specific rather than being due to the duration of distance of PACC exposure.

A Chi-Squared test was used to evaluate whether there was a significant association between the experimental population and a significant neighbour identity effect. There was no significant association between the experimental population and a significant neighbour identity effect at a 5% level. This means that within the PACC model a significant neighbour identity is independent of the experimental population. Hence further confirming that a significant neighbour identity effect is cell specific.

5.4.3 Invasion populations have increased morphological persistence

Finally, the level of morphological persistence was investigated among the 3 experimental populations. Across the 69 tracked cells 61% had a significant state matrix estimate that ranged from 0.16 to 0.98 where a smaller state matrix estimate signifies a higher level of morphological persistence. The remaining 39% of cells had a state matrix estimate that was not significantly different from 0. That is, the rate of morphological change at time t does not significantly depend on the rate of morphological change at time $t-1$. In turn, this suggests that an individual cell is more responsive because its morphological behaviour at time t is a

function of the current state dynamics, rather than also depending upon the previous value of the state.

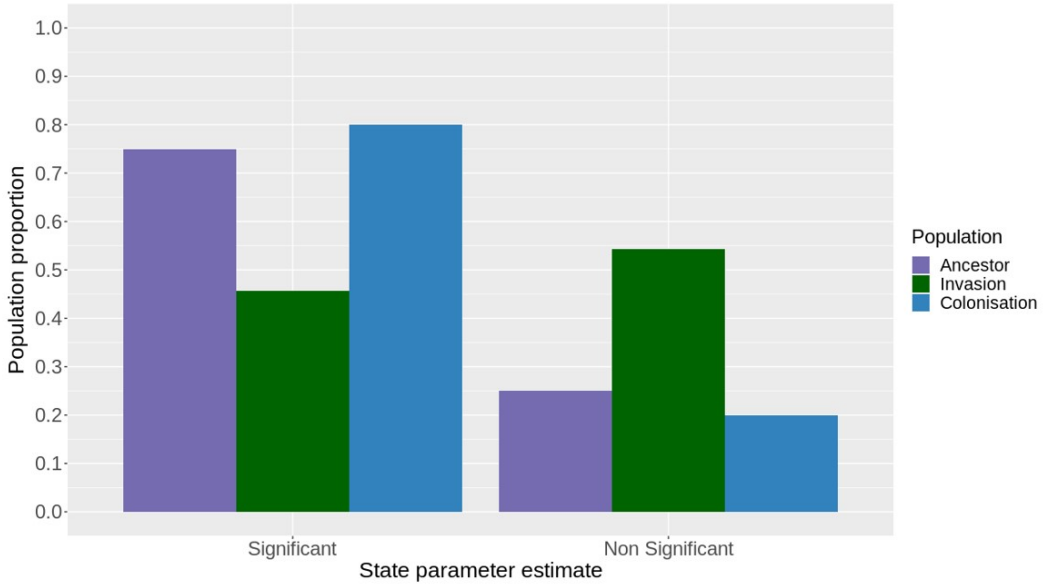


Figure 5.4: The proportion of cells within each population that have a significant state matrix estimate within the PACC model. A state matrix estimate is significant if the 95% confidence interval for that cell does not overlap 0. The cells are then stratified according to the significance of the state matrix estimate and the population type. The majority of cells within the ancestor and colonisation populations had a significant state estimate. In contrast, the majority of cells within the invasion populations had a non-significant state estimate. Hence this would suggest that the morphological behaviour of the invasion populations is more responsive due to the value at t being dependent on the current state of the system, rather than also depending on its own historical behaviour.

A Chi-Squared test was used to evaluate whether there was a significant association between the specific experimental population and the proportion of cells that have a significant temporally dependent structure (Figure 5.4). There was a significant association between the specific experimental population and the proportion of cells that have a significant temporally dependent structure at a 5% level ($p = 0.031$, $N = 69$). A post-hoc Bonferroni multiple comparison test was then used but it could not specify which populations were statistically different. This maybe due, in part, to the small overall sample size. Likewise, it may also be due to the invasion populations representing over half of the 69 tracked cells. Nevertheless,

the majority of cells within the ancestor and colonisation populations have a significant temporal structure in the rate of morphological change, 75% and 80% respectively. In contrast, the majority of cells within the invasion populations, 64%, have a non significant temporal structure in the rate of morphological change (Figure 5.4). Taken as a whole, this suggests that the invasion populations are more responsive to their current environment conditions compared to the ancestor or colonisation populations.

Finally, the 25 cells in which a neighbour identity effect was estimated were also investigated to determine whether there was a significant association between a significant state estimate and a significant PACC neighbour response. A Chi-Squared test found that there was no significant association between the significance of the state matrix estimate and a significant PACC neighbour response, hence reaffirming that a significant PACC response is cell specific.

5.5 Discussion

This chapter sought to investigate the migratory behaviour of PACCs to determine whether their increased size could prove to be problematic during metastatic dispersal. The chapter began by first comparing the rate of morphological and speed of migration between the PACC sub-population and the wider cancer cell population. The results showed the PACC sub-population is expected to have a higher rate of morphological change but a similar speed of migration. Hence these results supported the notion that whilst PACCs need more resources to survive, their migratory ability is not necessarily any better compared to a normal cancer cell. In turn, these results served as a foundation upon which to build a novel state space model to investigate whether PACCs interact with normal cancer cells as a form of cooperative search. That is, do normal cancer cells detect when their nearest neighbour is a PACC and significantly change their morphological behaviour as a result. The model found that there is a significant change in morphological behaviour in a subset of cells when their nearest neighbour is a PACC compared to a normal cancer cell. However, the same behaviour

is not observed in an alternative model when a cell is randomly chosen to be a "pseudo" PACC suggesting that the interaction is specific to the rare PACC sub-population.

5.5.1 An annoying obstacle or an attractive opportunity

The PACC model highlights that there is a significant change in morphological behaviour in a subset of cells when their neighbour is a PACC compared with a normal cancer cell. However, the model does not differentiate whether the significant response is due to a transfer of diffusible goods or whether the PACC is simply acting as a large obstacle to overcome. Separating these two competing hypotheses is challenging because PACCs are, by definition, extremely large. Thus, the alternative model would not necessarily capture this confounding effect. Likewise, due to the myriad of different diffusible products within the local vicinity it maybe hard to identify a specific PACC neighbour signal. Nevertheless, the model still provides critical insight and subsequent further knowledge maybe gained by stratifying the PACC population itself.

The term *PACC* is a broad label that includes both large mononuclear and multinuclear cells (Pienta et al., 2020). In both instances the individual cell has an increased amount of genomic material compared to a normal cancer cell. However, in the former the genomic material is contained within a single nucleus whereas in the latter there are multiple distinct nuclei. Evaluating whether the same significant change in morphological behaviour occurs irrespective of whether the PACC neighbour is a mono or multi nuclear cell may reveal the cause of the significant neighbour identity response.

For example, during migration cells are able to squeeze through gapes within the tightly packed ECM down to a threshold pore size that is equal to 10% of their cross sectional nuclear area (Wolf et al., 2013). Hence, within the PACC population there is expected to be a large difference in the minimum threshold pore size between mono and multi nuclear PACCs. If the cross sectional nuclear area of a mononuclear PACC is equal to x then the minimum pore size it can migrate through is equal to $\frac{x}{10}$. In contrast, a multinuclear PACC with the same cross sectional nuclear area, x, has a minimum pore size equal to $\frac{x}{10\times DN}$ were DN equals the number of distinct nuclei. This difference is based on the assumption that

each distinct nuclei has its own rate limiting size and thus the cell can be rotated such that all of the nuclei are complementary. As a result, a multinuclear PACC would need to perform fewer structural alterations within the local microenvironment compared to a mononuclear PACC. In turn, there would be less need for cooperation between a multinuclear PACC and a neighbouring cancer cell meaning that the identity effect may no longer be significant within the PACC model. Also, if the neighbouring cells are still attracted by the high concentration of diffusible products within the local multinuclear PACC vicinity, but the multinuclear PACCs are not in a symbiotic relationship, then the rate of morphological change should be higher within a multinuclear PACC compared to a mononuclear PACC at an equal neighbour cell density. This is because the multinuclear PACC needs to navigate through the crowded environment, based on the results in Chapter 4. Unfortunately, it was not possible to test this hypothesis with the current data due to the lack of specific nuclear staining.

5.5.2 Any decision is better than indecision

The significant identity covariate in the PACC model highlights that temporally dependent decisions are being made by individual cells in response to their own specific environmental conditions e.g. the presence of a neighbouring PACC. A similar temporally dependent decision process is also being made during the invasion of a distant site and this might explain why the majority of cells within the invasion population have a non-significant state estimate.

Invasion at a distant site begins with a cell becoming lodged within the tight confines of a narrow capillary (Stoletov et al., 2010). If the cell then detects that there is a high resource level within the local vicinity it will extravaste out of the blood and migrate into the foreign tissue (Mallin et al., 2020). The decision to re-initiate migration therefore requires an individual cell to be decisive and highly responsive to the current environmental conditions.

A similar selective pressure can be seen within the invasion assays where a cell must cross an environmental boundary to survive and reach the next round of selection. In contrast to the ancestor and colonisation assays, the invasion assay requires an individual cell to migrate across an environmental boundary between the 2D tissue culture plastic and the

3D Matrigel island. However, cancer cells are known to migrate quickly, and preferentially, along environmental boundaries (Keeton et al., 2018). Thus, the corresponding decision to invade must also be quick. As a result, an individual cell needs to be highly responsive to the current environmental conditions rather than relying upon its own historic exposure. Hence the morphological behaviour at time t is a function of the current state of the system rather than also depending upon the previous system state.

5.5.3 The challenges of capturing transient behaviours

In summary, capturing transient changes in morphological behaviour offers a unique insight into the decision process made by an individual cell during migration. The PACC model in this chapter identifies a previously unknown decision process in a subset of cells across all experimental populations and irrespective of the duration or distance away from a PACC. Whilst further work is still needed to determine whether the results in this chapter represent a symbiotic PACC - neighbour relationship, the results do highlight the importance of evaluating transient behaviours to reveal new cancer dynamics.

One of the major challenges to overcome whilst investigating PACC related behaviours is the low frequency at which PACCs appear within the migratory population. Admittedly, the time-lapse data in this thesis was not collected to specifically investigate PACC related behaviours. Nevertheless, across the 813 cells that were original tracked in Chapter 3 only 10 were found to be PACCs. Hence the PACC population represents approximately 1.2% of the experimentally evolved cells within this thesis. As a result, the sample size within this chapter is relatively small which in turn limits the depth of post-hoc analysis that can be performed. Similarly, the model could only be applied to 3 of the 4 experimentally evolved populations due to no PACCs being tracked within the escape populations. Whilst the sample size could be readily increased by recording more time-lapse videos this also causes the quantity of image data increase. In turn, automated tracking techniques will need to be further refined to ensure that the large data sets can be efficiently and accurately processed.

In addition to improving the post-hoc analysis an increased sample size would also allow for a more detailed primary PACC model to be built. The neighbour identity is currently

recorded as a binary variable in the PACC model. This means that the neighbour identity effect causes a sudden "jump" in the rate of morphological change. However, in reality, the neighbour identity effect is expected to cause a continuous gradual deviation in the rate of morphological change. Thus a more realistic model may have a nearest neighbour effect size that changes through time. Hence, once the model is fitted, the size of the nearest neighbour effect with and without a PACC neighbour can be compared. Furthermore, the improved model would also be able to capture whether the neighbour identity effect size changes through time. That is, does the presence of a PACC have a smaller effect on a cell that has already interacted with a PACC previously.

Chapter 6

Discussion

The spread of cancer cells from a primary tumour to distant sites around the body marks the deadliest stage of cancer progression, metastasis (Fares et al., 2020). Metastasis occurs in nearly all types of cancer and is leading cause of cancer related mortality (Pienta et al., 2020). Yet, metastasis is also widely considered to be the most inefficient stage in cancer progression. Only a small fraction of cells that initially leave the primary tumour are eventually successful in colonising at a distant site (Chaffer and Weinberg, 2011; Chambers et al., 2002). Identifying the traits and behaviours that separate the cells capable of successful metastasis is therefore central to combating metastatic disease. This thesis investigated the evolution of individual signal processing, through the proxy of morphological change, as a key determinate in metastatic success and a possible means by which to identify cells with increased metastatic potential.

6.1 Making sense of cancer cell migration

A myriad of different cell migration modes are observed during metastatic dispersal causing extensive variation to exist between individual cells (Friedl and Wolf, 2003, 2010). Likewise, due to the widespread spatial and temporal heterogeneity within the surrounding tumour micro-environment (Yuan, 2016) dispersal behaviours are also known to change dynamically during the migration of an individual cell (Butler et al., 2020). However, in spite of the enor-

mous complexity, migratory dynamics are commonly characterised by a single quantitative metric such as the average cell speed or turning angle (Meijering et al., 2009; Pincus and Theriot, 2007; Prasad and Alizadeh, 2018). As a result, this an oversimplified view of the underlying complexity and can fail to provide the maximum biological insight. In contrast, this thesis has shown in Chapters 3 - 5 that by adopted more sophisticated analytic approaches novel insight can be gained into the intricacies that exist during metastatic dispersal.

6.1.1 Dependent behaviours

Firstly, a central theme throughout this thesis has been the investigation of complex migratory behaviours that develop from the interaction between individual cellular traits. For example, when the rate of morphological change was evaluated in isolation the colonisation populations were indistinguishable from the other experimentally evolved populations. However, when the rate of morphological change was evaluated in response to the speed of migration and nearest neighbour distance, a wealth of behavioural variation was revealed (Chapter 3). Hence if extensive behavioural diversity can be seen from simple experimental selective regimes. Then the corresponding degree of behavioural diversity in vivo is expected to be staggering!

However, the full extent of the behavioural variation will not be understood without a change in analytic approach. That is, rather than quantifying and comparing the value of individual traits between populations e.g. the average cell speed or turning angle. The focus should instead be on characterising and quantifying the underlying behavioural mechanisms (Figure 6.1). In turn, this would enable a *Hallmarks of metastatic dispersal* to be formed that identifies common dispersal behaviours correlated with increased metastatic potential. Thus, whilst the magnitude of a given behaviour maybe population or cancer specific, akin to the difference in slopes between the invasion and escape populations (Figure 3.7), the fundamental characteristics should remain constant. A similar concept can be seen by the presence of chromosomal instability. That is, whilst the specific chromosomal aberrations that occur tend to vary between cancer types, the presence of increased chromosomal instability as a whole is indicative of cancer progression. As a result, a hallmarks of metastatic dispersal

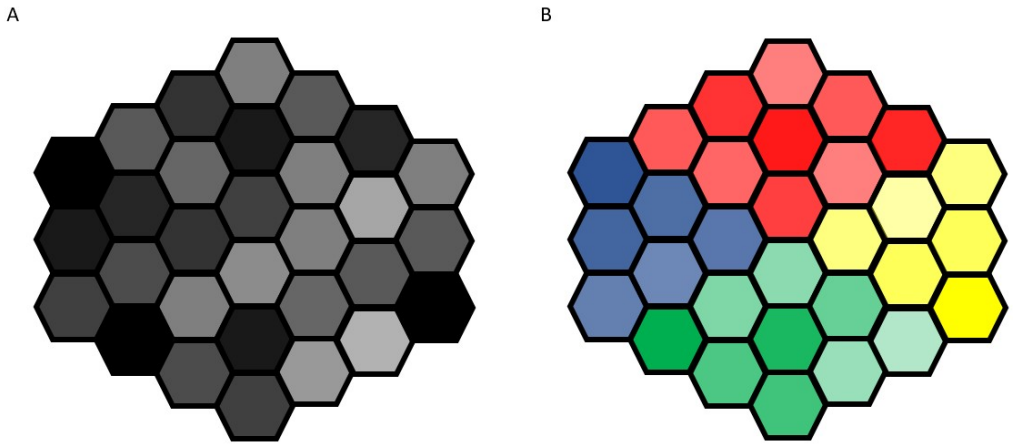


Figure 6.1: A graphical comparison of the difference in variation between isolated traits and dependent phenotypic behaviours. The shaded hexagons represent an individual cell where the colour and shade correspond to the behaviour and the intensity of the behaviour. (**A**) The grey scale represents a single migratory trait such as cell speed where the variation within the population is seen by the different shades. (**B**) The 4 main colours (red, yellow, green, and blue) represent different dependent behaviours that may exist within a population. The shade of each colour then signifies the average intensity of the given behaviour. Importantly, the dependent behaviours reveal a greater degree of variation within the population and also highlight distinct cellular sub-populations.

would then provide a natural link across multiple different cancers irrespective of the distant tissue type.

6.1.2 Temporal behaviours

Likewise, another core theme throughout this thesis has been the importance of characterising behaviours at a single cell level. For example, the temporally dependent morphological model in Chapter 4 showed that at a population level all of the experimental populations can adopt a similar complex morphological behaviour. However, at a single cell level, the same complex morphological behaviour was no longer present within the colonisation populations. Instead,

the single cell model showed that there were two similar, but yet different, morphological behaviours that existed together within the population.

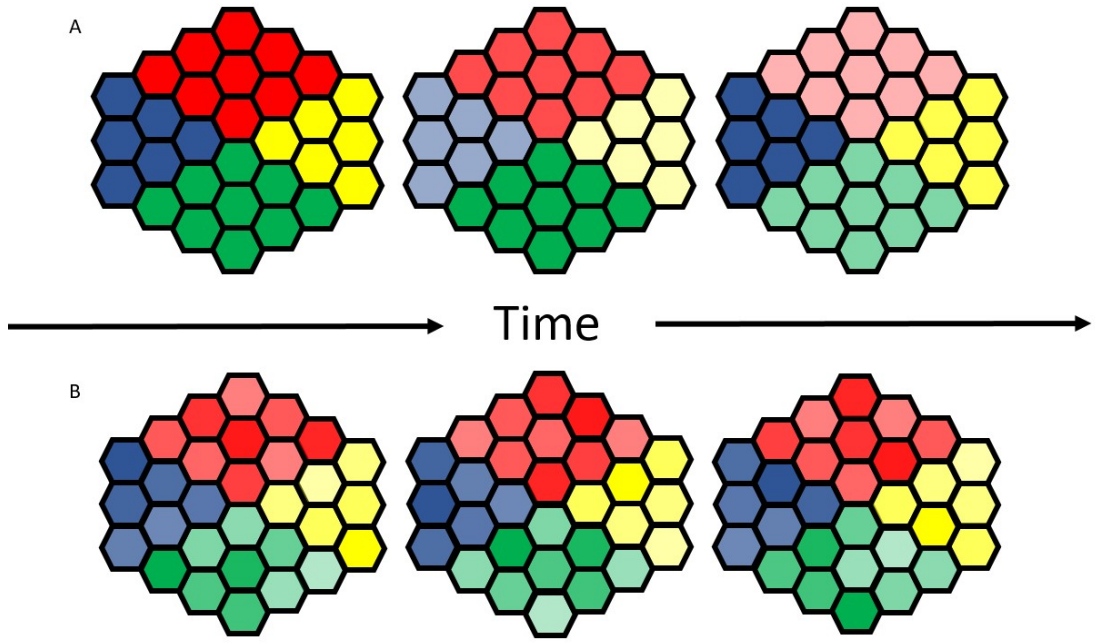


Figure 6.2: A graphical representation of the temporal changes in phenotypic behaviour at both a population and single cell level. The shaded hexagons represent an individual cell where the colour and shade correspond to the behaviour and the intensity of the behaviour. (**A**) The temporal variation in phenotypic behaviour at a population level as seen by the changes in shade of each colour across the distinct sub-populations. (**B**) The temporal variation in phenotypic behaviour at a single cell level as seen by the changes in shade of each individual cell. The single cell model captures the variation in time as well as the variation between individual cells. Hence the increased resolution enables the heterogeneity within the population to be quantified.

Characterising behaviours at a single cell level is important because it amalgamates multiple levels of biological variation. That is, the variation in migratory behaviour can be investigated together at a population, group, and single cell level (Figure 6.2). In the context of metastasis this is invaluable because whilst the majority of cells that leave a primary tumour will fail, metastatic success is achieved even if only a single cell is triumphant. Thus, gaining insight into the broader population level dynamic is important, but identifying the deadly subset of cells that are ultimately successful is also essential. Single cell phenotyping

therefore provides a framework in which to unravel the complex web of dynamics that develop at a both a population and single cell level during metastatic dispersal.

6.1.3 Transient behaviours

Finally, the penultimate chapter in this thesis (Chapter 5) focused on a specific sub-population of cells, poly-aneuploid cancer cells (PACCs). A temporally dependent morphological model found that a subset of cells displayed a significant, but transient, change in morphological behaviour when their nearest neighbour was a PACC compared with a normal cancer cell. In contrast, when a cell was chosen at random within the broader migratory population to be a "pseudo" PACC the same significant behaviour was not seen. Whilst multiple questions still exist regarding the specific role of PACCs in cancer evolution their presence offers an exciting opportunity to investigate a feature that maybe integral to lethal progression across multiple cancer types.

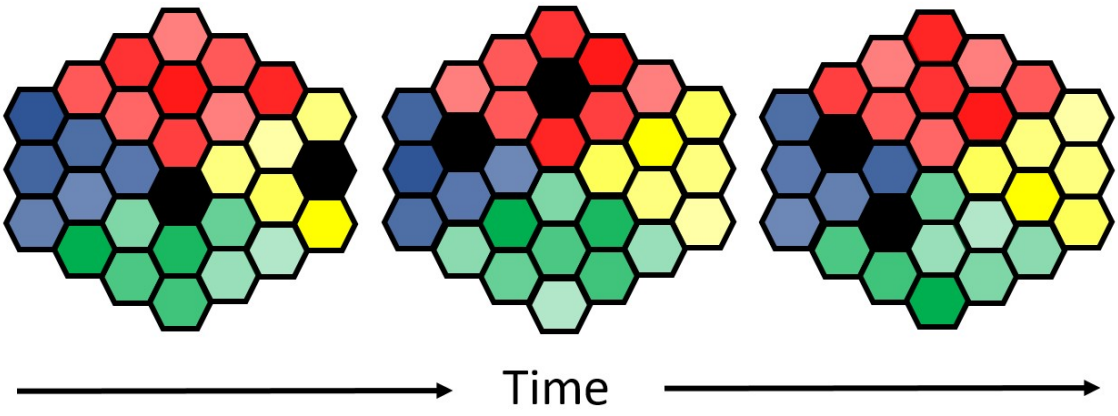


Figure 6.3: A graphical representation of a transient phenotypic behaviour.
The shaded hexagons represent an individual cell where the colour and shade correspond to the behaviour and the intensity of the behaviour. A black hexagon then represents a transient change in behaviour within the individual cell. As a result, the change in behaviour may then also have an effect on the spatially adjacent cells.

Yet the results in Chapter 5 also more broadly highlight the power of phenotypic analysis in capturing the specific spatial context. The importance of the surrounding microenvironment is widely acknowledged throughout cancer evolution (Yuan, 2016) and was a primary justification for the adoption of an experimental approach throughout this thesis. Yet, excluding the study of cellular quiescence, the effect of specific ecological disturbances on the behaviour of individual cells has been broadly ignored. Thus, in its most rudimentary form, the PACC model in Chapter 5 serves as a quantitative framework upon which further transient environmental effects can be investigated (Figure 6.3).

A further question regarding the formation of PACCs is whether the likelihood of cellular fusion between two specific cells is a function of the shared cellular ancestor? That is, do two "daughter" cells from a single division event detect the presence of one another and therefore possess a higher propensity for cellular fusion. Likewise, whilst a viable PACC may be more likely from the fusion of two "sister" cells, a PACC formed from the fusion of two ancestrally divergent cells may lead to a greater evolutionary burst of diversity. As a result, the identity of the constituent cells within a cell fusion derived PACC maybe as important as the PACC itself (Miroshnychenko et al., 2021). Yet, importantly, novel questions such as this can only be answered through a phenotypic approach in which the specific spatial context is retained on a single cell level.

6.2 Accounting for cell size in signal processing

The results in Chapter 5 also highlight a strong positive correlation between the rate of morphological change and cell area. That is, an increase in cellular area was seen to cause a corresponding increase in the rate of morphological change. However, due to the small sample size it was not possible to determine whether the correlation was present across all of the experimental populations or specific to the invasion population that accounted for over half of the cells within Chapter 5.

To test whether the affect of cell area on the rate of morphological change was population specific the same analytical approach from Chapter 3 was adopted with the addition of cell

area as another covariate. A linear mixed model was fitted across all of the data such that the rate of morphological change was dependent on the speed of migration, the distance to the nearest neighbouring cell, the interaction of the two, and the cell area. The model parameters were selected through a process of forward selection and only included if they were significant at the 5% level. The populations were also included as fixed effects allowing the intercepts and slope to vary between populations.

The cell area was found to be significantly and positively correlated with the rate of morphological change across all 4 of the experimental populations. That is, akin to the results in Chapter 5, an increase in cellular area was seen to cause a corresponding increase in the rate of morphological change. The inclusion of cellular area as a covariate also meant that the extended model explained a larger proportion of the variation compared with the previous model (marginal $R^2 = 0.609$ and 0.237 respectively).

In addition to cell area, the same significant covariate combinations were still present within the escape, invasion, and colonisation populations. That is, the cellular area and the speed of migration was significant within the invasion and escape populations whilst the full model was significant within the colonisation populations. However, a change in behaviour was found within the ancestor populations. In contrast to Chapter 3 where an intercept only model was fitted. The cell area and speed of migration were both found to be significantly and positively correlated with the rate of morphological change within the ancestor populations. Hence the ancestor populations can be seen to have the same morphological dynamics as both the escape and invasion population after accounting for the affect of cellular area.

In summary, signal processing is a complex behaviour that is dependent on multiple different cell extrinsic and intrinsic factors, one of which is cellular area. Hence whilst the affect of cell area on the signal processing behaviour of a cell offers an exciting area of research. The novel finding also underling the importance of capturing dependent behaviours when quantifying cellular dynamics during metastasis (Section 6.1.1). Finally, caution should always be taken when constructing complex mathematical models to ensure that all elements of the model accurately capture the underling biology of the system.

6.3 Challenges and limitations

A range of novel analytical approaches have been applied throughout this thesis to develop a deeper understanding of the migratory dynamics that exist during metastatic dispersal. However, a common obstacle that has plagued this thesis, as well as all forms of phenotypic analysis in cancer, is the challenge of obtaining accurate measures of phenotypic traits. This is acutely apparent when investigating complex behaviours such as signal processing that require large multi-dimensional data sets to be curated, manipulated, and then processed to make sense of the biological complexity that exists within. The challenges, limitations and possible solutions to some of the problems that were encountered during this thesis are discussed within the following section.

6.3.1 Experimental approach

Firstly, all of the phenotypic models within this thesis (Chapters 3 - 5) were built upon data obtained from 4 experimentally evolved populations of breast cancer cells. The populations were evolved using selective regimes that were designed to simulate different stages of the metastatic cascade (Figure 3.1).

One advantage of using an experimental approach is that populations of cells can be compared that only differ on their ability to metastasise (Sprouffske et al., 2012; Taylor et al., 2013). Hence the variation in phenotypic behaviours such as signal processing can be attributed to distinct selective pressures, an aspect that is not possible within a tumour. Yet, by the same reasoning, the experimental selective regimes may also be criticised for not accurately representing the complexity within the native system (Buckling et al., 2009). For example, the selective regimes in this thesis do not consider the transportation through the circulatory system, a crucial stage in metastatic dispersal. Likewise, the interaction between cancer cells and neighbouring stromal cells, a prominent feature during metastatic dispersal, is not captured. As a result, the findings in this thesis do not necessarily represent the dynamics that develop within a patient. However, the results do function as a proof of

concept and likewise all of the quantitative modelling approaches could be translated into another system.

An important future project will be to build upon the experimental approaches within this thesis and to apply the selective regimes in series as would be the case in vivo. Whilst this will lose the distinct selective pressures within each population it will validate whether the same colonisation dynamics are observed after earlier stages of experimentally selection. Likewise, identifying whether the same phenotypic behaviours appear suddenly within the escape and invasion populations after the first round of selection, or gradually after multiple rounds, will also provide value insight. In vivo multiple rounds of escape and invasion do not occur. Thus the selective regimes in this thesis may have unrealistically selecting populations of cells with lower levels of phenotypic variation.

6.3.2 Generating structured data

Next, whilst a plethora of different experimental protocols can be adopted the phenotypic behaviour of individual cells must still be recorded via a time-lapse video. In turn, the information that is encoded within the video, such as the location or shape of a cell, must be translated into a quantitative value that can be used for downstream analysis. This process is known as cell tracking and is a rate limiting step in many cell migration experiments.

In Chapter 2 a variety of different approaches were discussed to automate the cell tracking process. A convolutional neural network (CNN) was then used to automatically segment the morphology of 161,085 cells across 11,880 phase contrast time-lapse images. The segmented data then formed the basis for the subsequent morphological models presented in Chapters 3 - 5. Whilst a CNN was essential in this thesis there are still certain limitations that need to be address before wider penetration can be achieved within the cell migration community.

Firstly, in contrast to other automated tracking platforms such as ImageJ (Schindelin et al., 2012) or Ilastik (Sommer et al., 2011) CNNs do not typically possess a graphic user interface (GUI). Hence interacting within a CNN relies upon having a working knowledge of the command line. Whilst this knowledge can be quickly developed it may still perturb some experimental cell biologists from embracing the use of a CNN. Secondly, manual corrections

are often needed irrespective of how well the CNN has been trained. The data in this thesis required multiple weeks of manual curation to ensure that the segmented morphologies accurately matched the underlying ground truth. Admittedly, the segmentation performance will improve as the quantity of annotated training data increases and underling technology matures. Likewise, the degree of manual post-processing can be exponentially reduced in certain scenarios through the use of florescent tags. Nevertheless, the concept of fully automated cell tracking is not yet available and will most likely not be available in the foreseeable future. As a result, the expense of performing manual corrections needs to be carefully considered when designing experiments that require extensive cell tracking.

6.3.3 Analytical infrastructure

Finally, quantifying the migratory behaviour of individual cells requires the tracked trajectory of each cell to be post-processed to extract metrics such as the nearest neighbour distance. The post-processing stage was a major challenge within this thesis due to the lack of developed analytic infrastructure. As a result, a bespoke post-processing pipeline had to be developed from the ground up that was both efficient and scalable to handle the large quantity of data. This meant that advanced computational techniques were required that may not be within the remit of many experimental cell biologists. Hence if the field of cellular phenotyping is going to progress the need for skilled image-informaticians will be essential to ensure that suitable computational methods can be developed.

6.4 Future work

This thesis has shown that a variety of different migration behaviours can develop during metastatic dispersal. Yet, elucidating whether the same migratory dynamics emerge within a native 3D environmental remains unknown. The transition from quantifying cell migration in 2D to 3D is an important, but complex, leap that can reveal radically different migratory behaviours (Yamada and Sixt, 2019). However, more broadly, 3D phenotyping also provides an opportunity to integrate quantitative modelling directly into the experimental workflow.

This then preserves the natural population structure and also enables continuous changes in phenotypic behaviours to be captured on both a cellular and evolutionary timescale.

For example, in this thesis the migratory behaviour of the colonisation populations were recorded, and then quantified, at the end of the 6 month experiment on a 2D tissue culture plate. Hence this meant that only two evolution time-points were obtained and the cellular behaviours were recorded within a featureless environment. In contrast, by recording the cells within the native 3D tissue environment multiple phenotypic samples could have been collected over the 6 month duration. This would then enable the evolutionary trajectory of each population to "tracked" and understood directly, similar to the difference between quantifying phenotypic behaviours in time-lapse videos vs inferring behaviours from still images. As a result, a multi-level state space model could then estimate the evolutionary dynamics within each population from the individual migratory behaviours. In other words, multiple phenotypic snapshots could be taken, and the global cell population stratified by population type and migratory behaviour, akin to Figures 4.5 and 5.3. The snapshots could then be coalesced to form an "evolutionary time-lapse video" to detect whether convergent or divergent dynamics develop in response to specific selective regimes. In the context of poly-aneuploid cancer cells this could provide essential insight as the evolutionary trajectory of the population could be compared relative to the fluctuations in PACC frequency. That is, do increases in PACC frequency precede bursts in phenotypic diversity. If so, does the magnitude of the phenotypic burst decrease through time or do PACCs consistently provide a constant unwavering source of cellular diversity.

Furthermore, quantifying the migratory behaviours in situ also means that the environmental features and population structure can be retained. That is, the distinct cellular sub-populations will still be spatially partitioned. In the context of metastatic dispersal this enables questions regarding the influence of kin selection to be investigated because the migrant cell will be leaving a defined population structure. However, to fully benefit from the increased data richness a corresponding shift in the analytical approach will also be needed. Instead of measuring migratory rates, such as cell speed, the analytic focus will need to be on the raw geographic position of each cell, the migratory coordinates. Whilst this presents a

host of new quantitative challenges it also fully leverages the power of phenotypic analysis in capturing the spatial context in which a cellular dynamic develops.

References

Adler, C. E. and A. Sánchez Alvarado (2015, 11). Types or States? Cellular Dynamics and Regenerative Potential. *Trends in cell biology 25*(11), 687–696.

Aeberhard, W. H., J. Mills Flemming, and A. Nielsen (2018, 3). Review of State-Space Models for Fisheries Science. *Annual Review of Statistics and Its Application 5*(1), 215–235.

Aghdam, H. and E. Heravi (2017, 1). *Guide to Convolutional Neural Networks: A Practical Application to Traffic-Sign Detection and Classification*. Springer.

Aguilar, P. S., M. K. Baylies, A. Fleissner, L. Helming, N. Inoue, B. Podbilewicz, H. Wang, and M. Wong (2013, 7). Genetic basis of cell-cell fusion mechanisms. *Trends in genetics : TIG 29*(7), 427–437.

Alberts, B., J. H. Wilson, and T. Hunt (2008). *Molecular biology of the cell* (5 ed.). Garland Science.

Alizadeh, E., S. M. Lyons, J. M. Castle, and A. Prasad (2016). Measuring systematic changes in invasive cancer cell shape using Zernike moments. *Integr. Biol. 8*(11), 1183–1193.

Allan, D. B., T. Caswell, N. C. Keim, and C. M. van der Wel (2018). trackpy: Trackpy v0.4.1.

Amend, S. R. and A. J. Pienta (2015). Ecology meets cancer biology: the cancer swamp promotes the lethal cancer phenotype. *Oncotarget 6*(12), 9669–9678.

Amend, S. R., S. Roy, J. S. Brown, and K. J. Pienta (2016, 9). Ecological paradigms to understand the dynamics of metastasis. *Cancer Letters 380*(1), 237–242.

Ananthakrishnan, R. and A. Ehrlicher (2007, 6). The forces behind cell movement. *International journal of biological sciences 3*(5), 303–17.

Andor, N., T. A. Graham, M. Jansen, L. C. Xia, C. A. Aktipis, C. Petritsch, H. P. Ji, and C. C. Maley (2016, 1). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature medicine 22*(1), 105–113.

Anguiano, M., C. Castilla, M. Maška, C. Ederra, R. Peláez, X. Morales, G. Muñoz-Arrieta, M. Mujika, M. Kozubek, A. Muñoz-Barrutia, A. Rouzaut, S. Arana, J. M. Garcia-Aznar, and C. Ortiz-de Solorzano (2017, 2). Characterization of three-dimensional cancer cell migration in mixed collagen-Matrigel scaffolds using microfluidics and image analysis. *PLOS ONE 12*(2), e0171417.

Asgari Taghanaki, S., K. Abhishek, J. P. Cohen, J. Cohen-Adad, and G. Hamarneh (2020). Deep semantic segmentation of natural and medical images: a review. *Artificial Intelligence Review*.

Auger-Méthé, M., C. Field, C. M. Albertsen, A. E. Derocher, M. A. Lewis, I. D. Jonsen, and J. Mills Flemming (2016). State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. *Scientific Reports 6*(1), 26677.

Baltissen, D., T. Wollmann, M. Gunkel, I. Chung, H. Erfle, K. Rippe, and K. Rohr (2018). Comparison of segmentation methods for tissue microscopy images of glioblastoma cells. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 396–399.

Barton, N. H. (2007). *Evolution*. Cold Spring Harbor Laboratory Press.

Basanta, D. and A. R. A. Anderson (2017, 9). Homeostasis Back and Forth: An Ecoevolutionary Perspective of Cancer. *Cold Spring Harbor Perspectives in Medicine 7*(9), 1–23.

Ben-David, U. and A. Amon (2020). Context is everything: aneuploidy in cancer. *Nature Reviews Genetics 21*(1), 44–62.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning 2*(1), 1–127.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures.

Bengio, Y., A. Courville, and P. Vincent (2013, 8). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 35*(8), 1798–1828.

Berg, S., D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A. Hamprecht, and A. Kreshuk (2019). ilastik: interactive machine learning for (bio)image analysis. *Nature Methods 16*(12), 1226–1232.

Bernards, R. and R. A. Weinberg (2002, 8). Metastasis genes: A progression puzzle. *Nature 418*, 823.

Bielski, C. M., A. Zehir, A. V. Penson, M. T. A. Donoghue, W. Chatila, J. Armenia, M. T. Chang, A. M. Schram, P. Jonsson, C. Bandlamudi, P. Razavi, G. Iyer, M. E. Robson, Z. K. Stadler, N. Schultz, J. Baselga, D. B. Solit, D. M. Hyman, M. F. Berger, and B. S. Taylor (2018, 8). Genome doubling shapes the evolution and prognosis of advanced cancers. *Nature genetics 50*(8), 1189–1195.

Bollen, K. A. and R. W. Jackman (1985, 5). Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. *Sociological Methods & Research 13*(4), 510–542.

Bonte, D., H. Van Dyck, J. M. Bullock, A. Coulon, M. Delgado, M. Gibbs, V. Lehouck, E. Matthysen, K. Mustin, M. Saastamoinen, N. Schtickzelle, V. M. Stevens, S. Vandewoestijne, M. Baguette, K. Barton, T. G. Benton, A. Chaput-Bardy, J. Clobert, C. Dytham, T. Hovestadt, C. M. Meier, S. C. F. Palmer, C. Turlure, and J. M. J. Travis (2012, 5). Costs of dispersal. *Biological Reviews 87*(2), 290–312.

Bottou, L. and Y. Lecun (2004, 3). Large Scale Online Learning. *Advances in Neural Information Processing Systems 16*.

Bowler, D. E. and T. G. Benton (2005, 5). Causes and consequences of animal dispersal strategies: relating individual behaviour to spatial dynamics. *Biological Reviews 80*(2), 205–225.

Breier, G. (2000, 3). Angiogenesis in Embryonic Development—A Review. *Placenta 21*, S11–S15.

Bremnes, R. M., T. Dønnem, S. Al-Saad, K. Al-Shibli, S. Andersen, R. Sirera, C. Camps, I. Marinez, and L.-T. Busund (2011). The Role of Tumor Stroma in Cancer Progression and Prognosis: Emphasis on Carcinoma-Associated Fibroblasts and Non-small Cell Lung Cancer. *Journal of Thoracic Oncology 6*(1), 209–217.

Buckling, A., R. Craig Maclean, M. A. Brockhurst, and N. Colegrave (2009). The Beagle in a bottle. *Nature 457*(7231), 824–829.

Burrell, R. A., N. McGranahan, J. Bartek, and C. Swanton (2013). The causes and consequences of genetic heterogeneity in cancer evolution. *Nature 501*(7467), 338–345.

Butler, G., S. J. Keeton, L. J. Johnson, and P. R. Dash (2020, 12). A phenotypic switch in the dispersal strategy of breast cancer cells selected for metastatic colonization. *Proceedings. Biological sciences 287*(1940), 20202523.

Butler, G., J. Rudge, and P. R. Dash (2019, 10). Mathematical modelling of cell migration. *Essays in Biochemistry*.

Caicedo, J. C., S. Cooper, F. Heigwer, S. Warchal, P. Qiu, C. Molnar, A. S. Vasilevich, J. D. Barry, H. S. Bansal, O. Kraus, M. Wawer, L. Paavolainen, M. D. Herrmann, M. Rohban, J. Hung, H. Hennig, J. Concannon, I. Smith, P. A. Clemons, S. Singh, P. Rees, P. Horvath, R. G. Linington, and A. E. Carpenter (2017, 8). Data-analysis strategies for image-based cell profiling. *Nature Methods 14*, 849.

Calvi, B. R. (2013). Making big cells: One size does not fit all. *Proceedings of the National Academy of Sciences 110*(24), 9621–9622.

Cameron, L. A., P. A. Giardini, F. S. Soo, and J. A. Theriot (2000). Secrets of actin-based motility revealed by a bacterial pathogen. *Nature Reviews Molecular Cell Biology 1*(2), 110–119.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-8*(6), 679–698.

Caswell, D. R. and C. Swanton (2017). The role of tumour heterogeneity and clonal cooperativity in metastasis, immune evasion and clinical outcome. *BMC Medicine 15*(1), 133.

Chaffer, C. L. and R. A. Weinberg (2011, 3). A Perspective on Cancer Cell Metastasis. *Science 331*(6024), 1559 LP – 1564.

Chambers, A. F., A. C. Groom, and I. C. MacDonald (2002, 8). Dissemination and growth of cancer cells in metastatic sites. *Nature Reviews Cancer 2*, 563.

Chandra, A. and X. Yao (2006). Evolving hybrid ensembles of learning machines for better generalisation. *Neurocomputing 69*(7), 686–700.

Chen, J., K. Sprouffske, Q. Huang, and C. C. Maley (2011). Solving the Puzzle of Metastasis: The Evolution of Cell Migration in Neoplasms. *PLoS ONE 6*(4), e17933.

Clark, A. G. and D. M. Vignjevic (2015). Modes of cancer cell invasion and the role of the microenvironment. *Current Opinion in Cell Biology 36*, 13–22.

Clobert, J., M. Baguette, T. Benton, J. Bullock, and S. Ducatez (2012, 1). *Dispersal Ecology and Evolution*. Oxford University Press.

Clobert, J., J.-F. Le Galliard, J. Cote, S. Meylan, and M. Massot (2009, 3). Informed dispersal, heterogeneity in animal dispersal syndromes and the dynamics of spatially structured populations. *Ecology Letters 12*(3), 197–209.

Cooper, G. and R. E. Hausman (2000). *The cell ; a molecular approach*. Sinauer Associates.

Costa-Silva, B., N. M. Aiello, A. J. Ocean, S. Singh, H. Zhang, B. K. Thakur, A. Becker, A. Hoshino, M. T. Mark, H. Molina, J. Xiang, T. Zhang, T.-M. Theilen, G. García-Santos, C. Williams, Y. Ararso, Y. Huang, G. Rodrigues, T.-L. Shen, K. J. Labori, I. M. B. Lothe, E. H. Kure, J. Hernandez, A. Doussot, S. H. Ebbesen, P. M. Grandgenett, M. A. Hollingsworth, M. Jain, K. Mallya, S. K. Batra, W. R. Jarnagin, R. E. Schwartz, I. Matei, H. Peinado, B. Z. Stanger, J. Bromberg, and D. Lyden (2015, 6). Pancreatic cancer exosomes initiate pre-metastatic niche formation in the liver. *Nature cell biology 17*(6), 816–826.

Coward, J. and A. Harding (2014). Size Does Matter: Why Polyploid Tumor Cells are Critical Drug Targets in the War on Cancer. *Frontiers in Oncology 4*, 123.

Cowden Dahl, K. D., R. Dahl, J. N. Kruichak, and L. G. Hudson (2009, 11). The epidermal growth factor receptor responsive miR-125a represses mesenchymal morphology in ovarian cancer cells. *Neoplasia (New York, N.Y.) 11*(11), 1208–1215.

Crocker, J. C. and D. G. Grier (1996). Methods of Digital Video Microscopy for Colloidal Studies. *Journal of Colloid and Interface Science 179*(1), 298–310.

Dagogo-Jack, I. and A. T. Shaw (2018). Tumour heterogeneity and resistance to cancer therapies. *Nature Reviews Clinical Oncology 15*(2), 81–94.

Deryugina, E. I. and W. B. Kiosses (2017, 4). Intratumoral Cancer Cell Intravasation Can Occur Independent of Invasion into the Adjacent Stroma. *Cell reports 19*(3), 601–616.

Donovan, P., K. Cato, R. Legaie, R. Jayalath, G. Olsson, B. Hall, S. Olson, S. Boros, B. A. Reynolds, and A. Harding (2014, 4). Hyperdiploid tumor cells increase phenotypic heterogeneity within Glioblastoma tumors. *Molecular bioSystems 10*(4), 741–758.

Duchi, J., E. Hazan, and Y. Singer (2011, 7). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research 12*, 2121–2159.

Durbin, J. and S. J. Koopman (2001, 8). *Time Series Analysis by State Space Methods*. Oxford University Press.

Elena, S. F. and R. E. Lenski (2003). Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nature Reviews Genetics 4*(6), 457–469.

Erenpreisa, J., K. Salmina, A. Huna, E. A. Kosmacek, M. S. Cragg, F. Ianzini, and A. P. Anisimov (2011, 7). Polyploid tumour cells elicit paradiploid progeny through depolyploidizing divisions and regulated autophagic degradation. *Cell biology international 35*(7), 687–695.

Erickson, B. J., P. Korfiatis, Z. Akkus, and T. L. Kline (2017). Machine Learning for Medical Imaging. *Radiographics : a review publication of the Radiological Society of North America, Inc 37*(2), 505–515.

Everingham, M., S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2015). The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision 111*(1), 98–136.

Fan, B., Y. Zhou, Q. Ma, Q. Yu, C. Zhao, and K. Sun (2018). The Bet-Hedging Strategies for Seedling Emergence of Calligonum mongolicum to Adapt to the Extreme Desert Environments in Northwestern China .

Fang, M., J. Yuan, C. Peng, and Y. Li (2014, 4). Collagen as a double-edged sword in tumor progression. *Tumour Biology 35*(4), 2871–2882.

Fares, J., M. Y. Fares, H. H. Khachfe, H. A. Salhab, and Y. Fares (2020). Molecular principles of metastasis: a hallmark of cancer revisited. *Signal Transduction and Targeted Therapy 5*(1), 28.

Flusser, J., T. Suk, and B. Zitova (2016). *2D and 3D image analysis by moments*. Wiley-Blackwell.

Flusser, J., B. Zitova, and T. Suk (2009). *Moments and Moment Invariants in Pattern Recognition*. New York, UNITED KINGDOM: John Wiley & Sons, Incorporated.

Fortunato, A., A. Boddy, D. Mallo, A. Aktipis, C. C. Maley, and J. W. Pepper (2017, 2). Natural Selection in Cancer Biology: From Molecular Snowflakes to Trait Hallmarks. *Cold Spring Harbor perspectives in medicine 7*(2).

Fox, D. T., D. E. Soltis, P. S. Soltis, T.-L. Ashman, and Y. Van de Peer (2020). Polyploidy: A Biological Force From Cells to Ecosystems. *Trends in Cell Biology 30*(9), 688–694.

Friedl, P., J. Locker, E. Sahai, and J. E. Segall (2012, 8). Classifying collective cancer cell invasion. *Nature Cell Biology 14*, 777.

Friedl, P. and K. Wolf (2003, 5). Tumour-cell invasion and migration: diversity and escape mechanisms. *Nature Reviews Cancer 3*, 362.

Friedl, P. and K. Wolf (2010, 1). Plasticity of cell migration: a multiscale tuning model. *The Journal of Cell Biology 188*(1), 11–19.

Futuyma, D. J. and G. Moreno (1988, 11). THE EVOLUTION OF ECOLOGICAL SPECIALIZATION. *Annual Review of Ecology and Systematics 19*(1), 207–233.

Gaggioli, C., S. Hooper, C. Hidalgo-Carcedo, R. Grosse, J. F. Marshall, K. Harrington, and E. Sahai (2007). Fibroblast-led collective invasion of carcinoma cells with differing roles for RhoGTPases in leading and following cells. *Nature Cell Biology 9*(12), 1392–1400.

Gardner, M. W. and S. R. Dorling (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment 32*(14), 2627–2636.

Gerrish, P. J. and P. D. Sniegowski (2012, 9). Real time forecasting of near-future evolution. *Journal of the Royal Society, Interface 9*(74), 2268–2278.

Ghalambor, C. K., J. K. McKay, S. P. Carroll, and D. N. Reznick (2007). Adaptive Versus Non-Adaptive Phenotypic Plasticity and the Potential for Contemporary Adaptation in New Environments. *Functional Ecology 21*(3), 394–407.

Giancotti, F. G. (2013, 11). Mechanisms governing metastatic dormancy and reactivation. *Cell 155*(4), 750–764.

Gillies, R. J., J. S. Brown, A. R. A. Anderson, and R. A. Gatenby (2018). Eco-evolutionary causes and consequences of temporal changes in intratumoural blood flow. *Nature Reviews Cancer 18*(9), 576–585.

Girshick, R., J. Donahue, T. Darrell, and J. Malik (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, Washington, DC, USA, pp. 580–587. IEEE Computer Society.

Gkretsi, V. and T. Stylianopoulos (2018, 5). Cell Adhesion and Matrix Stiffness: Coordinating Cancer Cell Invasion and Metastasis. *Frontiers in oncology 8*, 145.

Glorot, X., A. Bordes, and Y. B. B. T. P. o. t. F. I. C. o. A. I. a. Statistics (2011, 6). Deep Sparse Rectifier Neural Networks.

Goldberg, A. D., C. D. Allis, and E. Bernstein (2007). Epigenetics: A Landscape Takes Shape. *Cell 128*(4), 635–638.

Gomis, R. R. and S. Gawrzak (2016). Tumor cell dormancy. *Molecular oncology 11*(1), 62.

Goodfellow Ian, Bengio Yoshua, and Courville Aaron (2016). *Deep Learning*. MIT Press.

Gordon, D. J., B. Resio, and D. Pellman (2012). Causes and consequences of aneuploidy in cancer. *Nature Reviews Genetics 13*(3), 189–203.

Gordonov, S., M. K. Hwang, A. Wells, F. B. Gertler, D. A. Lauffenburger, and M. Bathe (2016, 1). Time series modeling of live-cell shape dynamics for image-based phenotypic profiling. *Integrative biology : quantitative biosciences from nano to macro 8*(1), 73–90.

Graham, B. (2014, 12). Fractional Max-Pooling.

Graham, T. A. and A. Sottoriva (2017, 1). Measuring cancer evolution from the genome. *The Journal of Pathology 241*(2), 183–191.

Greaves, M. and C. C. Maley (2012, 1). Clonal evolution in cancer. *Nature 481*, 306.

Grigore, A. D., M. K. Jolly, D. Jia, M. C. Farach-Carson, and H. Levine (2016, 4). Tumor Budding: The Name is EMT. Partial EMT. *Journal of clinical medicine 5*(5), 1–23.

Gupta, G. P. and J. Massagué (2006). Cancer Metastasis: Building a Framework. *Cell 127*(4), 679–695.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. (Parts I and II). *Journal of Theoretical Biology 7*(1), 1–16.

Hamilton, W. D. and R. M. May (1977). Dispersal in stable habitats. *Nature 269*(5629), 578–581.

Hanahan, D. and R. A. Weinberg (2000, 1). The hallmarks of cancer. *Cell 100*(1), 57–70.

Hanahan, D. and R. A. Weinberg (2011). Hallmarks of cancer: the next generation. *Cell 144*, 1–29.

Harvey, A. C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.

He, K., G. Gkioxari, P. Dollár, and R. Girshick (2017). Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 2980–2988.

He, K., X. Zhang, S. Ren, and J. Sun (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778.

Holmes, E. E., E. J. Ward, and K. Wills (2012). MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *The R Journal 4*(1), 30.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE Transactions on Information Theory 8*(2), 179–187.

Huang, J., V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy (2016, 11). Speed/accuracy trade-offs for modern convolutional object detectors. *CoRR*.

Huth, J., M. Buchholz, J. M. Kraus, M. Schmucker, G. von Wichert, D. Krndija, T. Seufferlein, T. M. Gress, and H. A. Kestler (2010). Significantly improved precision of cell migration analysis in time-lapse video microscopy through use of a fully automated tracking system. *BMC Cell Biology 11*(1), 24.

Kalluri, R. and R. A. Weinberg (2009, 6). The basics of epithelial-mesenchymal transition. *The Journal of Clinical Investigation 119*(6), 1420–1428.

Kalman, R. E. (1960, 3). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering 82*(1), 35–45.

Katt, M. E., A. D. Wong, and P. C. Searson (2018). Dissemination from a Solid Tumor: Examining the Multiple Parallel Pathways. *Trends in Cancer 4*(1), 20–37.

Kawecki, T. J., R. E. Lenski, D. Ebert, B. Hollis, I. Olivieri, and M. C. Whitlock (2012). Experimental evolution. *Trends in Ecology & Evolution 27*(10), 547–560.

Kayalibay, B., G. Jensen, and P. van der Smagt (2017). CNN-based Segmentation of Medical Imaging Data.

Keeton, S. J., J. M. Delalande, M. Cranfield, A. Burns, and P. R. Dash (2018). Compressed collagen and decellularized tissue – novel components in a pipeline approach for the study of cancer metastasis. *BMC Cancer 18*(1), 622.

Keskar, N. S. and R. Socher (2017). Improving Generalization Performance by Switching from Adam to SGD.

Khan, A., A. Sohail, U. Zahoora, and A. S. Qureshi (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*.

Kim, J., J. K. Lee, and K. M. Lee (2016). Accurate Image Super-Resolution Using Very Deep Convolutional Networks.

Kim, S. H., J. Turnbull, and S. Guimond (2011). Extracellular matrix and cell signalling: The dynamic cooperation of integrin, proteoglycan and growth factor receptor.

Kimmel, J. C., A. Y. Chang, A. S. Brack, and W. F. Marshall (2018, 1). Inferring cell state by quantitative motility analysis reveals a dynamic state system and broken detailed balance. *PLoS computational biology 14*(1), e1005927.

King, J. G. and J. D. Hadfield (2019, 2). The evolution of phenotypic plasticity when environments fluctuate in time and space. *Evolution Letters 3*(1), 15–27.

Kingma, D. P. and J. Ba (2014). Adam: A Method for Stochastic Optimization. *CoRR abs/1412.6980*, 1–15.

Klein, C. A. (2009). Parallel progression of primary tumours and metastases. *Nature Reviews Cancer 9*(4), 302–312.

Kohn, J. C., M. C. Lampi, and C. A. Reinhart-King (2015, 3). Age-related vascular stiffening: causes and consequences. *Frontiers in genetics 6*, 112.

Kondor, R. and S. Trivedi (2018, 2). On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups. *CoRR*.

Kotsiantis, S. B., I. D. Zaharakis, and P. E. Pintelas (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review 26*(3), 159–190.

Krizhevsky, A., I. Sutskever, and G. Hinton (2012, 1). ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems 25*.

Lambert, A. W., D. R. Pattabiraman, and R. A. Weinberg (2017). Emerging Biological Principles of Metastasis. *Cell 168*(4), 670–691.

Lauffenburger, D. A. and A. F. Horwitz (1996). Cell Migration: A Physically Integrated Molecular Process. *Cell 84*(3), 359–369.

Lawrence, S., C. L. Giles, A. C. Tsoi, and A. D. Back (1997). Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks 8*(1), 98–113.

LeCun, Y., Y. Bengio, and G. Hinton (2015). Deep learning. *Nature 521*(7553), 436–444.

Lewontin, R. C. (1970, 11). The Units of Selection. *Annual Review of Ecology and Systematics 1*(1), 1–18.

Li, H., Z. Xu, G. Taylor, C. Studer, and T. Goldstein (2018). Visualizing the Loss Landscape of Neural Nets. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31*, pp. 6389–6399. Curran Associates, Inc.

Liao, S. X. (1994). *Image Analysis by Moments*. Ph. D. thesis, The University of Manitoba.

Liao, S. X. and M. Pawlak (1996). On image analysis by moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence 18*(3), 254–266.

Lin, K.-C., G. Torga, Y. Sun, R. Axelrod, K. J. Pienta, J. C. Sturm, and R. H. Austin (2019, 4). The role of heterogeneous environment and docetaxel gradient in the emergence of polyploid, mesenchymal and resistant prostate cancer cells. *Clinical & experimental metastasis 36*(2), 97–108.

Lin, T., P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie (2017). Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944.

Lin, T.-Y., M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár (2015). Microsoft COCO: Common Objects in Context.

Liu, H.-S., M.-S. Jan, C.-K. Chou, P.-H. Chen, and N.-J. Ke (1999). Is Green Fluorescent Protein Toxic to the Living Cells? *Biochemical and Biophysical Research Communications 260*(3), 712–717.

López, S., E. L. Lim, S. Horswell, K. Haase, A. Huebner, M. Dietzen, T. P. Mourikis, T. B. K. Watkins, A. Rowan, S. M. Dewhurst, N. J. Birkbak, G. A. Wilson, P. Van Loo, M. Jamal-Hanjani, T. Consortium, C. Swanton, and N. McGranahan (2020, 3). Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nature genetics 52*(3), 283–293.

Lopez-Sánchez, L. M., C. Jimenez, A. Valverde, V. Hernandez, J. Peñarando, A. Martinez, C. Lopez-Pedrera, J. R. Muñoz-Castañeda, J. R. De la Haba-Rodríguez, E. Aranda, and A. Rodriguez-Ariza (2014, 6). CoCl2, a Mimic of Hypoxia, Induces Formation of Polyploid Giant Cells with Stem Characteristics in Colon Cancer. *PLOS ONE 9*(6), e99143.

Ltkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Springer Publishing Company, Incorporated.

Luzzi, K. J., I. C. MacDonald, E. E. Schmidt, N. Kerkvliet, V. L. Morris, A. F. Chambers, and A. C. Groom (1998). Multistep nature of metastatic inefficiency: dormancy of solitary cells after successful extravasation and limited survival of early micrometastases. *The American journal of pathology 153*(3), 865–873.

Lyons, S. M., E. Alizadeh, J. Mannheimer, K. Schuamberg, J. Castle, B. Schroder, P. Turk, D. Thamm, and A. Prasad (2016, 3). Changes in cell shape are correlated with metastatic potential in murine and human osteosarcomas. *Biology Open 5*(3), 289 LP – 299.

Malanchi, I., A. Santamaria-Martínez, E. Susanto, H. Peng, H.-A. Lehr, J.-F. Delaloye, and J. Huelsken (2012). Interactions between cancer stem cells and their niche govern metastatic colonization. *Nature 481*(7379), 85–89.

Maley, C. C., P. C. Galipeau, J. C. Finley, V. J. Wongsurawat, X. Li, C. A. Sanchez, T. G. Paulson, P. L. Blount, R.-A. Risques, P. S. Rabinovitch, and B. J. Reid (2006). Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nature Genetics 38*(4), 468–473.

Maley, C. C. and B. J. Reid (2005). Natural selection in neoplastic progression of Barrett's esophagus. *Seminars in Cancer Biology 15*(6), 474–483.

Mallin, M. M., K. J. Pienta, and S. R. Amend (2020). Cancer cell foraging to explain bone-specific metastatic progression. *Bone*, 115788.

Mannan, R., X. Wang, P. S. Bawa, D. E. Spratt, A. Wilson, J. Jentzen, A. M. Chinnaiyan, Z. R. Reichert, and R. Mehra (2020, 2). Polypoidal giant cancer cells in metastatic castration-resistant prostate cancer: observations from the Michigan Legacy Tissue Program. *Medical oncology (Northwood, London, England) 37*(3), 16.

Marshall, W. F., K. D. Young, M. Swaffer, E. Wood, P. Nurse, A. Kimura, J. Frankel, J. Wallingford, V. Walbot, X. Qu, and A. H. K. Roeder (2012). What determines cell size? *BMC Biology 10*(1), 101.

Martincorena, I., K. M. Raine, M. Gerstung, K. J. Dawson, K. Haase, P. Van Loo, H. Davies, M. R. Stratton, and P. J. Campbell (2017). Universal patterns of selection in cancer and somatic tissues. *Cell 171*.

Marusyk, A. and K. Polyak (2010, 1). Tumor heterogeneity: causes and consequences. *Biochimica et biophysica acta 1805*(1), 105–117.

Massagué, J. and A. C. Obenauf (2016, 1). Metastatic colonization by circulating tumour cells. *Nature 529*(7586), 298–306.

Masuzzo, P., M. Van Troys, C. Ampe, and L. Martens (2016, 2). Taking Aim at Moving Targets in Computational Cell Migration. *Trends in Cell Biology 26*(2), 88–110.

Matthysen, E. (2005, 6). Density-dependent dispersal in birds and mammals. *Ecography 28*(3), 403–416.

McGranahan, N. and C. Swanton (2015). Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell 27*.

McGranahan, N. and C. Swanton (2017). Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell 168*(4), 613–628.

McPeek, M. A. and R. D. Holt (1992). The Evolution of Dispersal in Spatially and Temporally Varying Environments. *The American Naturalist 140*(6), 1010–1027.

Medberry, C. J., P. M. Crapo, B. F. Siu, C. A. Carruthers, M. T. Wolf, S. P. Nagarkar, V. Agrawal, K. E. Jones, J. Kelly, S. A. Johnson, S. S. Velankar, S. C. Watkins, M. Modo, and S. F. Badylak (2013, 1). Hydrogels derived from central nervous system extracellular matrix. *Biomaterials 34*(4), 1033–1040.

Meijering, E., O. Dzyubachyk, and I. Smal (2012). Chapter nine - Methods for Cell and Particle Tracking. In P. M. B. T. M. i. E. conn (Ed.), *Imaging and Spectroscopic Analysis of Living Cells*, Volume 504, pp. 183–200. Academic Press.

Meijering, E., O. Dzyubachyk, I. Smal, and W. A. van Cappellen (2009). Tracking in cell and developmental biology. *Seminars in Cell & Developmental Biology 20*(8), 894–902.

Merlo, L. M. F., J. W. Pepper, B. J. Reid, and C. C. Maley (2006, 11). Cancer as an evolutionary and ecological process. *Nature Reviews Cancer 6*, 924.

Meyer, A. S. and L. M. Heiser (2019). Systems biology approaches to measure and model phenotypic heterogeneity in cancer. *Current Opinion in Systems Biology 17*, 35–40.

Michod, R. E. (1999). *Darwinian dynamics: evolutionary transitions in fitness and individuality*. Princeton: Princeton University Press.

Miroshnychenko, D., E. Baratchart, M. C. Ferrall-Fairbanks, R. V. Velde, M. A. Laurie, M. M. Bui, A. C. Tan, P. M. Altrock, D. Basanta, and A. Marusyk (2021). Spontaneous cell fusions as a mechanism of parasexual recombination in tumour cell populations. *Nature Ecology & Evolution 5*(3), 379–391.

Moen, E., D. Bannon, T. Kudo, W. Graf, M. Covert, and D. Van Valen (2019). Deep learning for cellular image analysis. *Nature Methods 16*(12), 1233–1246.

Morris, L. G. T., N. Riaz, A. Desrichard, Y. Şenbabaoğlu, A. A. Hakimi, V. Makarov, J. S. Reis-Filho, and T. A. Chan (2016). Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget; Vol 7, No 9*.

Murlidhar, V., R. M. Reddy, S. Fouladdel, L. Zhao, M. K. Ishikawa, S. Grabauskiene, Z. Zhang, J. Lin, A. C. Chang, P. Carrott, W. R. Lynch, M. B. Orringer, C. Kumar-Sinha, N. Palanisamy, D. G. Beer, M. S. Wicha, N. Ramnath, E. Azizi, and S. Nagrath (2017, 9). Poor Prognosis Indicated by Venous Circulating Tumor Cell Clusters in Early-Stage Lung Cancers. *Cancer research 77*(18), 5194–5206.

Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

Nakagawa, S. and H. Schielzeth (2013, 2). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution 4*(2), 133–142.

Nakajima, E. C. and B. Van Houten (2013, 5). Metabolic symbiosis in cancer: Refocusing the Warburg lens. *Molecular Carcinogenesis 52*(5), 329–337.

Nixon, M. and A. S. Aguado (2012). *Feature Extraction & Image Processing for Computer Vision, Third Edition* (3rd ed.). USA: Academic Press, Inc.

Nowell, P. C. (1976, 10). The clonal evolution of tumor cell populations. *Science 194*(4260), 23 LP – 28.

Odenwald, M. A., J. R. Prosperi, and K. H. Goss (2013, 1). APC/β-catenin-rich complexes at membrane protrusions regulate mammary tumor cell migration and mesenchymal morphology. *BMC cancer 13*, 12.

Olgac, A. and B. Karlik (2011, 2). Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks. *International Journal of Artificial Intelligence And Expert Systems 1*, 111–122.

Olson, M. F. and E. Sahai (2008). The actin cytoskeleton in cancer cell motility. *Clinical & Experimental Metastasis 26*(4), 273.

Orr, H. A. (2009, 8). Fitness and its role in evolutionary genetics. *Nature reviews. Genetics 10*(8), 531–539.

Orr-Weaver, T. L. (2015, 6). When bigger is better: the role of polyploidy in organogenesis. *Trends in genetics : TIG 31*(6), 307–315.

Page-McCaw, A., A. J. Ewald, and Z. Werb (2007, 3). Matrix metalloproteinases and the regulation of tissue remodelling.

Paget, S. (1889). The distribution of secondary growths in cancer of the breast. *The Lancet 133*(3421), 571–573.

Pan, S. J. and Q. Yang (2010). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering 22*(10), 1345–1359.

Pandya, P., J. L. Orgaz, and V. Sanz-Moreno (2017). Actomyosin contractility and collective migration: may the force be with you.

Pantel, K. and R. H. Brakenhoff (2004, 6). Dissecting the metastatic cascade. *Nature Reviews Cancer 4*(6), 448–456.

Parker, T. M., V. Henriques, A. Beltran, H. Nakshatri, and R. Gogna (2020). Cell competition and tumor heterogeneity. *Seminars in Cancer Biology 63*, 1–10.

Paul, C. D., W.-C. Hung, D. Wirtz, and K. Konstantopoulos (2016, 7). Engineered Models of Confined Cell Migration. *Annual review of biomedical engineering 18*, 159–80.

Pavel, M., M. Renna, S. J. Park, F. M. Menzies, T. Ricketts, J. Füllgrabe, A. Ashkenazi, R. A. Frake, A. C. Lombarte, C. F. Bento, K. Franze, and D. C. Rubinsztein (2018). Contact inhibition controls cell survival and proliferation via YAP/TAZ-autophagy axis. *Nature Communications 9*(1), 2961.

Pavlides, S., D. Whitaker-Menezes, R. Castello-Cros, N. Flomenberg, A. K. Witkiewicz, P. G. Frank, M. C. Casimiro, C. Wang, P. Fortina, S. Addya, R. G. Pestell, U. E. Martinez-Outschoorn, F. Sotgia, and M. P. Lisanti (2009, 12). The reverse Warburg effect: aerobic glycolysis in cancer associated fibroblasts and the tumor stroma. *Cell cycle (Georgetown, Tex.) 8*(23), 3984–4001.

Peinado, H., S. Lavotshkin, and D. Lyden (2011). The secreted factors responsible for pre-metastatic niche formation: Old sayings and new thoughts. *Seminars in Cancer Biology 21*(2), 139–146.

Petrie, R. J. and K. M. Yamada (2012, 12). At the leading edge of three-dimensional cell migration. *Journal of cell science 125*(Pt 24), 5917–26.

Pienta, K. J., E. U. Hammarlund, R. H. Austin, R. Axelrod, J. S. Brown, and S. R. Amend (2020). Cancer cells employ an evolutionarily conserved polyploidization program to resist therapy. *Seminars in Cancer Biology*.

Pienta, K. J., E. U. Hammarlund, R. Axelrod, S. R. Amend, and J. S. Brown (2020, 6). Convergent Evolution, Evolving Evolvability, and the Origins of Lethal Cancer. *Molecular cancer research : MCR 18*(6), 801–810.

Pienta, K. J., E. U. Hammarlund, R. Axelrod, J. S. Brown, and S. R. Amend (2020, 2). Poly-aneuploid cancer cells promote evolvability, generating lethal cancer. *Evolutionary Applications n/a*(n/a).

Pienta, K. J., E. U. Hammarlund, J. S. Brown, S. R. Amend, and R. M. Axelrod (2021). Cancer recurrence and lethality are enabled by enhanced survival and reversible cell cycle arrest of polyaneuploid cells. *Proceedings of the National Academy of Sciences 118*(7).

Pincus, Z. and J. A. Theriot (2007, 9). Comparison of quantitative methods for cell-shape analysis. *Journal of Microscopy 227*(2), 140–156.

Polyak, K. (2007). Breast cancer: origins and evolution. *The Journal of clinical investigation 117*(11), 3155–3163.

Prasad, A. and E. Alizadeh (2018, 12). Cell Form and Function: Interpreting and Controlling the Shape of Adherent Cells. *Trends in Biotechnology*, 11.

Pries, A. R., B. Reglin, and T. W. Secomb (2001, 9). Structural adaptation of microvascular networks: functional roles of adaptive responses. *American journal of physiology. Heart and circulatory physiology 281*(3), 1015–25.

Provenzano, P. P., D. R. Inman, K. W. Eliceiri, J. G. Knittel, L. Yan, C. T. Rueden, J. G. White, and P. J. Keely (2008). Collagen density promotes mammary tumor initiation and progression. *BMC medicine 6*, 11.

Psaila, B., R. N. Kaplan, E. R. Port, and D. Lyden. Priming the 'soil' for breast cancer metastasis: the pre-metastatic niche. *Breast disease 26*, 65–74.

Puig, P.-E., M.-N. Guilly, A. Bouchot, N. Droin, D. Cathelin, F. Bouyer, L. Favier, F. Ghiringhelli, G. Kroemer, E. Solary, F. Martin, and B. Chauffert (2008, 9). Tumor cells can escape DNA-damaging cisplatin through DNA endoreduplication and reversible polyploidy. *Cell biology international 32*(9), 1031–1043.

Quatromoni, J. G., S. Singhal, P. Bhojnagarwala, W. W. Hancock, S. M. Albelda, and E. Eruslanov (2015, 1). An optimized disaggregation method for human lung tumors that preserves the phenotype and function of the immune cells. *Journal of Leukocyte Biology 97*(1), 201–209.

Raica, M., A. M. Cimpean, and D. Ribatti (2009, 7). Angiogenesis in pre-malignant conditions. *European Journal of Cancer 45*(11), 1924–1934.

Rangamani, P., A. Lipshtat, E. Azeloglu, R. Calizo, M. Hu, S. Ghassemi, J. Hone, S. Scarlata, S. Neves, and R. Iyengar (2013). Decoding Information in Cell Shape. *Cell 154*(6), 1356–1369.

Rees, M., C. Jessica, C. J. Metcalf, and D. Childs (2009, 12). Bet-hedging as an evolutionary game: The trade-off between egg size and number. *Proceedings. Biological sciences / The Royal Society 277*, 1149–1151.

Ren, S., K. He, R. Girshick, and J. Sun (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems 28*, pp. 91–99. Curran Associates, Inc.

Rezatofighi, H., N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese (2019, 2). Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *CoRR abs/1902.0*.

Rieger, H. and M. Welter (2015, 5). Integrative models of vascular remodeling during tumor growth. *Wiley interdisciplinary reviews. Systems biology and medicine 7*(3), 113–129.

Roeder, A. H. K., A. Cunha, M. C. Burl, and E. M. Meyerowitz (2012, 9). A computational image analysis glossary for biologists. *Development 139*(17), 3071.

Roweis, S. and Z. Ghahramani (1999). A Unifying Review of Linear Gaussian Models. *Neural Computation 11*(2), 305–345.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning representations by back-propagating errors. *Nature 323*(6088), 533–536.

Sainath, T. N., A. Mohamed, B. Kingsbury, and B. Ramabhadran (2013). Deep convolutional neural networks for LVCSR. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8614–8618.

Sansregret, L. and C. Swanton (2017, 1). The Role of Aneuploidy in Cancer Evolution. *Cold Spring Harbor perspectives in medicine 7*(1), a028373.

Sceneay, J., M. J. Smyth, and A. Möller (2013, 12). The pre-metastatic niche: finding common ground. *Cancer metastasis reviews 32*(3-4), 449–464.

Schindelin, J., I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Tinevez, D. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona (2012, 6). Fiji: an open-source platform for biological-image analysis. *Nature methods 9*(7), 676–682.

Sezgin, M. and B. Sankur (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic imaging 13*, 13–20.

Sharma, S. V., D. Y. Lee, B. Li, M. P. Quinlan, F. Takahashi, S. Maheswaran, U. McDermott, N. Azizian, L. Zou, M. A. Fischbach, K.-K. Wong, K. Brandstetter, B. Wittner, S. Ramaswamy, M. Classon, and J. Settleman (2010, 4). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. *Cell 141*(1), 69–80.

Shieh, A. C. (2011). Biomechanical Forces Shape the Tumor Microenvironment. *Annals of Biomedical Engineering 39*(5), 1379–1389.

Shumway, R. and D. Stoffer (2011, 1). *Time Series Analysis and Its Applications With R Examples*, Volume 9. Springer.

Shumway, R. H. and D. S. Stoffer (2017). *Time Series Analysis and Its Applications*. Springer International Publishing AG.

Shutler, J. D. and M. S. Nixon (2006). Zernike velocity moments for sequence-based description of moving features. *Image and Vision Computing 24*(4), 343–356.

Simonyan, K. and A. Zisserman (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition.

Siple, M. C. and T. B. Francis (2016). Population diversity in Pacific herring of the Puget Sound, USA. *Oecologia 180*(1), 111–125.

Smith, J. M. (1964, 3). Group Selection and Kin Selection. *Nature 201*, 1145.

Sommer, C., C. Straehle, U. Köthe, and F. A. Hamprecht (2011). Ilastik: Interactive learning and segmentation toolkit. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 230–233.

Sottoriva, A., H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press, and D. Shibata (2015). A Big Bang model of human colorectal tumor growth. *Nat Genet 47*.

Sprouffske, K., L. M. F. Merlo, P. J. Gerrish, C. C. Maley, and P. D. Sniegowski (2012). Cancer in light of experimental evolution.

Staneva, R., F. El Marjou, J. Barbazan, D. Krndija, S. Richon, A. G. Clark, and D. M. Vignjevic (2019, 3). Cancer cells in the tumor core exhibit spatially coordinated migration patterns. *Journal of Cell Science 132*(6), jcs220277.

Stoletov, K., H. Kato, E. Zardouzian, J. Kelber, J. Yang, S. Shattil, and R. Klemke (2010, 7). Visualizing extravasation dynamics of metastatic tumor cells. *Journal of Cell Science 123*(13), 2332 LP – 2341.

Svensson, C.-M., A. Medyukhina, I. Belyaev, N. Al-Zaben, and M. T. Figge (2018, 3). Untangling cell tracks: Quantifying cell migration by time lapse image data analysis. *Cytometry Part A 93*(3), 357–370.

Tahmasbi, A., F. Saki, and S. B. Shokouhi (2011). Classification of benign and malignant masses based on Zernike moments. *Computers in Biology and Medicine 41*(8), 726–735.

Tam, W. L. and R. A. Weinberg (2013, 11). The epigenetics of epithelial-mesenchymal plasticity in cancer. *Nature Medicine 19*(11), 1438–1449.

Tay, S., J. J. Hughey, T. K. Lee, T. Lipniacki, S. R. Quake, and M. W. Covert (2010, 7). Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature 466*(7303), 267–271.

Taylor, A. M., J. Shih, G. Ha, G. F. Gao, X. Zhang, A. C. Berger, S. E. Schumacher, C. Wang, H. Hu, J. Liu, A. J. Lazar, C. G. A. R. Network, A. D. Cherniack, R. Beroukhim, and M. Meyerson (2018, 4). Genomic and Functional Approaches to Understanding Cancer Aneuploidy. *Cancer cell 33*(4), 676–689.

Taylor, T. B., L. J. Johnson, R. W. Jackson, M. A. Brockhurst, and P. R. Dash (2013, 4). First steps in experimental cancer evolution. *Evolutionary Applications 6*(3), 535–548.

Taylor, T. B., A. V. Wass, L. J. Johnson, and P. Dash (2017, 12). Resource competition promotes tumour expansion in experimentally evolved cancer. *BMC Evolutionary Biology 17*(1), 268.

Teh, C. . and R. T. Chin (1988). On image analysis by the methods of moments. *IEEE Transactions on Pattern Analysis and Machine Intelligence 10*(4), 496–513.

Théry, M. and M. Bornens (2006). Cell shape and cell division. *Current Opinion in Cell Biology 18*(6), 648–657.

Thygesen, U. H., C. M. Albertsen, C. W. Berg, K. Kristensen, and A. Nielsen (2017). Validation of ecological state space models using the Laplace approximation. *Environmental and Ecological Statistics 24*(2), 317–339.

Tissot, T., F. Massol, B. Ujvari, C. Alix-Panabieres, N. Loeuille, and F. Thomas (2019, 12). Metastasis and the evolution of dispersal. *Proceedings of the Royal Society B: Biological Sciences 286*(1916), 20192186.

Tonnesen, M. G., X. Feng, and R. A. Clark (2000, 12). Angiogenesis in Wound Healing. *Journal of Investigative Dermatology Symposium Proceedings 5*(1), 40–46.

Tsai, H. F., J. Gajda, T. F. Sloan, A. Rares, and A. Q. Shen (2019, 1). Usiigaci: Instance-aware cell tracking in stain-free phase contrast microscopy enabled by machine learning. *SoftwareX 9*, 230–237.

Tumer, K. and J. Ghosh (1996). Estimating the Bayes error rate through classifier combining. In *Proceedings of 13th International Conference on Pattern Recognition*, Volume 2, pp. 695–699.

Turajlic, S., A. Sottoriva, T. Graham, and C. Swanton (2019). Resolving genetic heterogeneity in cancer. *Nature Reviews Genetics 20*(7), 404–416.

Turajlic, S. and C. Swanton (2016, 4). Metastasis as an evolutionary process. *Science 352*(6282), 169 LP – 175.

Tweedy, L., B. Meier, J. Stephan, D. Heinrich, and R. G. Endres (2013). Distinct cell shapes determine accurate chemotaxis. *Scientific Reports 3*(1), 2606.

Valastyan, S. and R. A. Weinberg (2011, 10). Tumor metastasis: molecular insights and evolving paradigms. *Cell 147*(2), 275–92.

Van Valen, D. A., T. Kudo, K. M. Lane, D. N. Macklin, N. T. Quach, M. M. DeFelice, I. Maayan, Y. Tanouchi, E. A. Ashley, and M. W. Covert (2016). Deep Learning Automates the Quantitative Analysis of Individual Cells in Live-Cell Imaging Experiments. *PLOS Computational Biology 12*(11), e1005177.

Vazquez-Martin, A., O. V. Anatskaya, A. Giuliani, J. Erenpreisa, S. Huang, K. Salmina, I. Inashkina, A. Huna, N. N. Nikolsky, and A. E. Vinogradov (2016, 11). Somatic polyploidy is associated with the upregulation of c-MYC interacting genes and EMT-like signature. *Oncotarget 7*(46), 75235–75260.

Venkatesan, S. and C. Swanton (2016, 5). Tumor Evolutionary Principles: How Intratumor Heterogeneity Influences Cancer Treatment and Outcome. *American Society of Clinical Oncology Educational Book* (36), e141–e149.

Vicar, T., J. Balvan, J. Jaros, F. Jug, R. Kolar, M. Masarik, and J. Gumulec (2019). Cell segmentation methods for label-free contrast microscopy: review and comprehensive comparison. *BMC Bioinformatics 20*(1), 360.

Villa Martín, P., M. A. Muñoz, and S. Pigolotti (2019, 4). Bet-hedging strategies in expanding populations. *PLOS Computational Biology 15*(4), e1006529.

Wang, Z., J. D. Butner, R. Kerketta, V. Cristini, and T. S. Deisboeck (2015, 2). Simulating cancer growth with multiscale agent-based modeling. *Seminars in cancer biology 30*, 70–8.

Webb, A. R., K. D. Copsey, and G. Cawley (2011). *Statistical Pattern Recognition*. Hoboken, UNITED KINGDOM: John Wiley & Sons, Incorporated.

Weinberg, R. A. (1996). How Cancer Arises. *Scientific American 275*, 62–70.

Weinberg, R. A. (2007). *The biology of cancer*. Garland Science.

Wells, A., J. Grahovac, S. Wheeler, B. Ma, and D. Lauffenburger (2013, 5). Targeting tumor cell motility as a strategy against invasion and metastasis. *Trends in pharmacological sciences 34*(5), 283–289.

Wershof, E., D. Park, R. P. Jenkins, D. J. Barry, E. Sahai, and P. A. Bates (2019, 10). Matrix feedback enables diverse higher-order patterning of the extracellular matrix. *PLOS Computational Biology 15*(10), e1007251.

Wolf, K., M. Te Lindert, M. Krause, S. Alexander, J. Te Riet, A. L. Willis, R. M. Hoffman, C. G. Figdor, S. J. Weiss, and P. Friedl (2013, 6). Physical limits of cell migration: control by ECM space and nuclear deformation and tuning by proteolysis and traction force. *The Journal of cell biology 201*(7), 1069–84.

Wright, S. (1922). Coefficients of Inbreeding and Relationship. *The American Naturalist 56*(645), 330–338.

Wu, P.-H., D. M. Gilkes, J. M. Phillip, A. Narkar, T. W.-T. Cheng, J. Marchand, M.-H. Lee, R. Li, and D. Wirtz (2020, 1). Single-cell morphology encodes metastatic potential. *Science Advances 6*(4), eaaw6938.

Yamada, K. M. and M. Sixt (2019). Mechanisms of 3D cell migration. *Nature Reviews Molecular Cell Biology*.

Yao, G. (2014, 6). Modelling mammalian cellular quiescence. *Interface focus 4*(3), 20130074.

Yi, S., S. Lin, Y. Li, W. Zhao, G. B. Mills, and N. Sahni (2017). Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nature Reviews Genetics 18*(7), 395–410.

Yuan, Y. (2016, 8). Spatial Heterogeneity in the Tumor Microenvironment. *Cold Spring Harbor perspectives in medicine 6*(8), a026583.

Zeiler, M. (2012, 12). ADADELTA: An adaptive learning rate method. *CoRR 1212*.

Zeiler, M. D. and R. Fergus (2013). Visualizing and Understanding Convolutional Networks. *CoRR abs/1311.2*.

Zernike, F. (1942). Phase contrast, a new method for the microscopic observation of transparent objects part II. *Physica 9*(10), 974–986.

Zernike, F. and F. J. M. Stratton (1934). Diffraction Theory of the Knife-Edge Test and its Improved Form, The Phase-Contrast Method. *Monthly Notices of the Royal Astronomical Society 94*(5), 377–384.

Zhang, J., K. F. Goliwas, W. Wang, P. V. Taufalele, F. Bordeleau, and C. A. Reinhart-King (2019, 4). Energetic regulation of coordinated leader–follower dynamics during collective invasion of breast cancer cells. *Proceedings of the National Academy of Sciences 116*(16), 7867 LP – 7872.

Zhang, W., R. Li, T. Zeng, Q. Sun, S. Kumar, J. Ye, and S. Ji (2020, 6). Deep Model Based Transfer and Multi-Task Learning for Biological Image Analysis. *IEEE Transactions on Big Data 6*(2), 322–333.

Zhao, Z., P. Zheng, S. Xu, and X. Wu (2019). Object Detection With Deep Learning: A Review. *IEEE Transactions on Neural Networks and Learning Systems 30*(11), 3212–3232.

Zimmet, J. and K. Ravid (2000). Polyploidy: Occurrence in nature, mechanisms, and significance for the megakaryocyte-platelet system. *Experimental Hematology 28*(1), 3–16.

Zuur, A., I. Tuck, and N. Bailey (2003, 5). Dynamic factor analysis to estimate common trends in fisheries time series. *Canadian Journal of Fisheries and Aquatic Sciences - CAN J FISHERIES AQUAT SCI 60*, 542–552.

Zuur, A. F., R. J. Fryer, I. T. Jolliffe, R. Dekker, and J. J. Beukema (2003). Estimating common trends in multivariate time series using dynamic factor analysis. *Environmetrics 14*(7), 665–685.

Appendix A

Experimental methods

A.1 Escape Assay

Initially, MDA-MB-231 cells (LGC) were encapsulated in a 2mg/ml collagen gel (rat-tail collagen type 1, First Link) and set into a 24-well plate which was used as a mould (750,000 cells per gel, Greiner Bio-One). The collagen gels were compressed for 2 minutes as described in Keeton et al. (2018), then set into a 1mg/ml low density collagen gel (rat tail collagen type 1, First Link). Once set, cell culture medium (Dulbeco's Modified Eagles Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS), and Penicillin 100 µg/ml, Streptomycin 100 U/ml (Gibco, Fisher Scientific)) was added over the top. Medium was replaced every 3-4 days. After 10-14 days, the compressed collagen disc was separated from the low density collagen and collagenase type 1 diluted in phospho-buffered saline solution (Gibco, Fisher Scientific) used to retrieve the cells from the collagen matrix, 200 U/ml for compressed collagen and 100 U/ml for low density collagen. Cells in collagenase/PBS were incubated at 37°C in a stirred water-bath at 45 rpm for 30-60 minutes, then washed in Phospo-buffered saline solution (PBS, Gibco Fisher Scientific). Cells extracted from the compressed collagen were placed in liquid nitrogen storage and those collected from the low density collagen were seeded into 2mg/ml collagen gel with medium over for population expansion. Once expanded, cells were retrieved from collagen

using collagenase in PBS then seeded into 2mg/ml collagen for compression or frozen at -80°C and transferred to liquid nitrogen for storage.

A.2 Invasion Assay

MDA-MB-231 cells (LGC) were re-suspended in PBS, and seeded around the outside of a 5mg/ml set Matrigel island in a 6-well plate Matrigel (#35623, Corning), was diluted using DMEM without supplements. Cells were seeded in excess at the island margins, with around 40,000 cells seeded in 200µl per experiment for the initial set-up. Cells were left to settle and adhere to the 2D surface for 60 minutes then cell culture medium added over the top (DMEM supplemented with 10% FBS, and penicillin 100 µg/ml, streptomycin 100 U/ml). Medium was changed every 3-4 days and cells were harvested after 7 days. Cells were retrieved from Matrigel using Cell Recovery solution (#354253, Corning) on ice for 45-60 minutes, washed with ice cold PBS then reseeded into Matrigel at 5mg/ml to expand cell numbers. After 7 days the cells were released from Matrigel using cell recovery solution as described above (typically 400,000 – 500,000 per gel), re-suspended in PBS and seeded in excess around the outside of a new Matrigel island (5mg/ml) for the next round of the 2D/3D invasion assay or cells were frozen at -80°C and transferred to liquid nitrogen for storage.

A.3 Colonisation Assay

Rat lung was retrieved from 9 week old Wistar rats (Envigo) and flash frozen. It was then thawed and decellularized using repeated rounds of treatment following an adapted version of the protocol published in Medberry et al. (2013). Briefly: frozen lung was thawed and cut into small pieces of around 100mg, which were then placed into deionized water (ddH2O), stirred at 60 rpm for 16 hours at 4°C. Lung tissue was treated with 0.02% trypsin/0.05% EDTA for 60 minutes at 37°C at 60 rpm, 3% Triton-X 100/PBS for 70 minutes, 1M sucrose/PBS for 30 minutes, 4% deoxycholate/ddH2O for 60 minutes, 0.1% peracetic acid in 4% ethanol for 120 minutes, PBS for 5 minutes, and finally twice in ddH2O for 15 minutes. The

tissue was washed thoroughly between each treatment with ddH2O. De-cellularization was checked between rounds using epifluorescence microscopy and staining with DAPI H1200 Vectashield (Vectorlabs) to identify whether cell nuclei remained within the matrix structure. Decellularized lung tissue was freeze-dried and stored in an airtight container.

Using decellularized lung as a culture matrix: tissue was soaked in 70% ethanol, washed with PBS and then rehydrated in PBS pH 7.2 (Gibco) in a tissue culture incubator for 5 days, then soaked in cell culture medium (DMEM supplemented with 10% FBS and penicillin/streptomycin as described above) for 48 hours. Cells grown in 2D tissue culture flasks were trypsinized, re-suspended in medium then 750,000 cells added in low volume of medium (100-150 µl) over the decellularized lung tissue in a 6-well plate and left to adhere for 2 hours. Medium was then added over the top so that the decellularized lung rafts floated. Rafts were transferred to new wells when the bottom of the well was confluent with shed and adhered cells. To feed the cells growing in/on the raft, ½ of the medium (2ml of 4ml) was aspirated and replaced every 2-4 days. After 140 and 189 days, rafts were retrieved from medium, washed with PBS and cells harvested by incubating in: collagenase I (170 U/ml, Gibco 17018-029), collagenase IV (170 U/ml, Gibco 17104-019), elastase (0.075 U/ml, Sigma E7885) (based on the protocol described in Quatromoni et al. (2015)) incubated at 37°C 45rpm in a stirred water-bath, then washed twice with PBS before seeding in 2D tissue culture plates for expansion. Expanded cells were then frozen at -80°C and transferred to liquid nitrogen for storage.

A.4 Time-lapse microscopy

The frozen cells were retrieved from liquid nitrogen, cultured in 2D tissue culture flasks (25cm^2 or 75cm^2 Greiner bio-one), trypsinized and seeded into 6-well plates (Greiner bio-one) at 10-15% cell confluence. Time-lapse movies were made for 12 hour periods with images taken at 2 minute intervals, using a Nikon TiE phase contrast microscope with an environmental chamber (37°C) and moveable platform stage. x10 Plan Apo DIC L Lens

was used in conjunction with an intermediate magnification changer set to x1.5 to give x15 magnification. NIS Elements software was used for image capture.

Appendix B

Supplementary figures

B.1 Chapter 3

Population	Mean rate of morphological change	SE rate of morphological change	Mean speed of migration	SE speed of migration	N
Ancestor	4.782	0.074	4.466	0.096	88
Escape	4.900	0.037	4.238	0.045	230
Invasion	4.717	0.033	4.234	0.040	283
Colonisation	4.822	0.038	4.310	0.045	212

Table B.1: The natural log-transformed mean and standard error for the rate of morphological change and the speed of migration. Displayed are the natural log mean and standard error for the rate of morphological change and speed of migration for each of the four populations. The escape populations have a significantly higher rate of morphological change compared with the invasion populations, p = 0.0289.

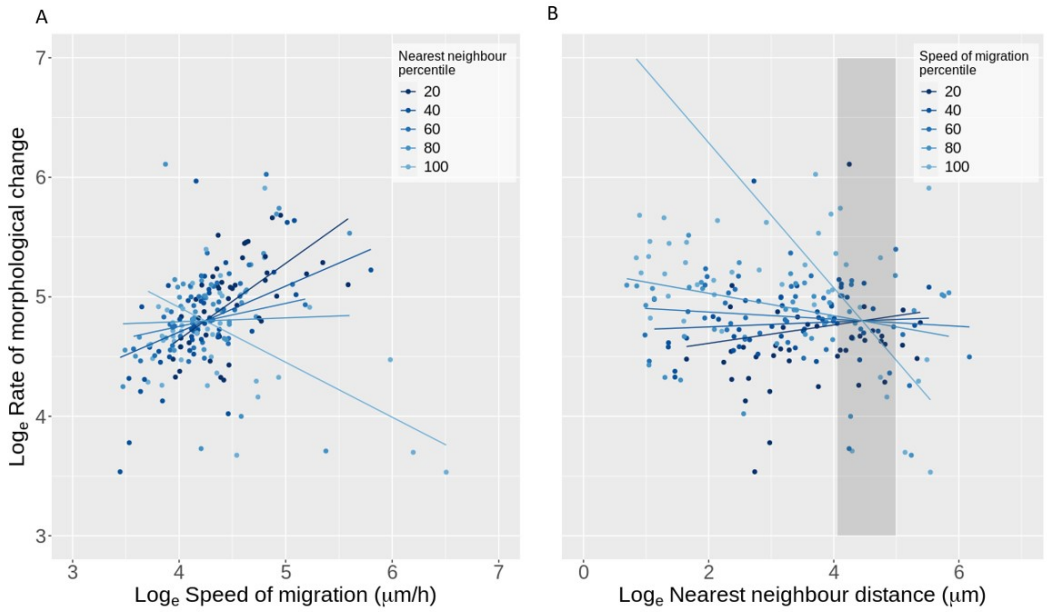


Figure B.1: A dynamic switch in the morphological behaviour within cells selected for colonisation with data points. The speed of migration (p = 5.418 x10-14), the distance to the nearest neighbouring cell (p = 2.207 x10-10) and the interaction of the two (2.219 x10-11) was significant in the colonisation population (N = 210). (A) The natural log-transformed rate of morphological change against the natural log-transformed speed of migration. The data point colour relates to the distance from a neighbouring cell. The lighter the data point the further away from a neighbouring cell. The shaded lines represent the predicted natural log-transformed rate of morphological change against the natural log-transformed speed of migration. The shaded lines indicate the natural log-transformed nearest neighbour percentile. The light the line the further away from a neighbouring cell. (B) The natural log-transformed rate of morphological change against the natural log-transformed nearest neighbour distance. The data point colour relates to the speed of migration. The lighter the data point the faster the speed of migration. The shaded lines represent the predicted natural log-transformed rate of morphological change against the natural log-transformed nearest neighbour distance. The shaded lines indicate the speed of migration percentile. The lighter the line the faster the speed of migration. The shaded region indicates the range of distances over which there is no significant relationship in the rate of morphological change and the speed of migration when the data is centred at these distances, between 57.9μm and 147.2μm.

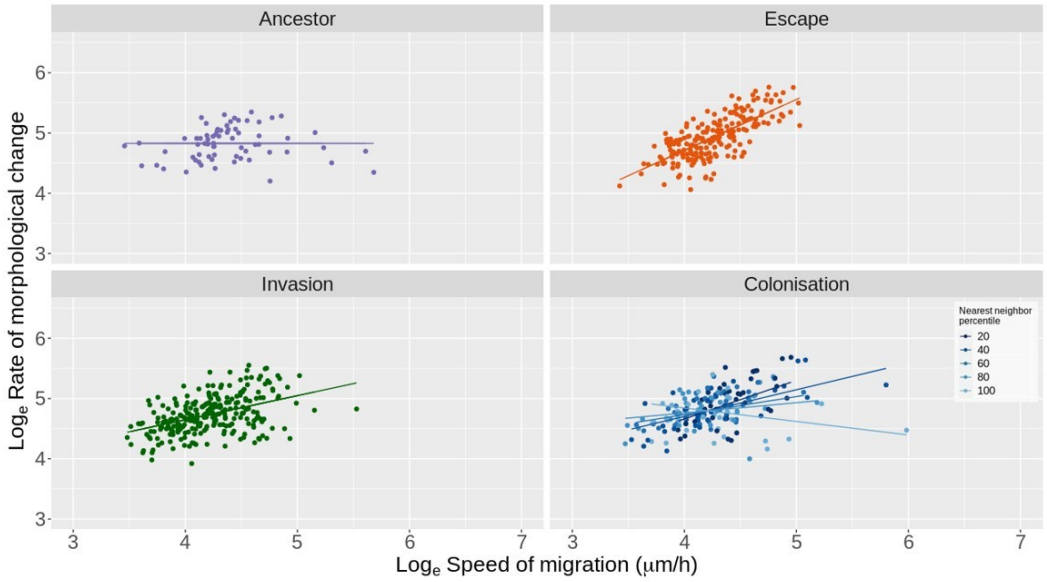


Figure B.2: The reduced model for each population after the removal of influential data points. The natural log-transformed rate of morphological change against the natural log-transformed speed of migration. In the colonisation populations the shaded lines indicate the natural log-transformed nearest neighbour percentile. The lighter the line the further away from a neighbouring cell. Influential data points, Cook's distance > (4 / N) where N is the sample size (Bollen and Jackman, 1985), have been removed to test whether a small subset of points influencing the result. After the removal of the influential points the speed of migration was still significant in the escape and invasion populations. Likewise, the speed of migration, distance to the nearest neighbouring cell and the interaction of the two was still significant in the colonisation populations.

B.2 Chapter 4

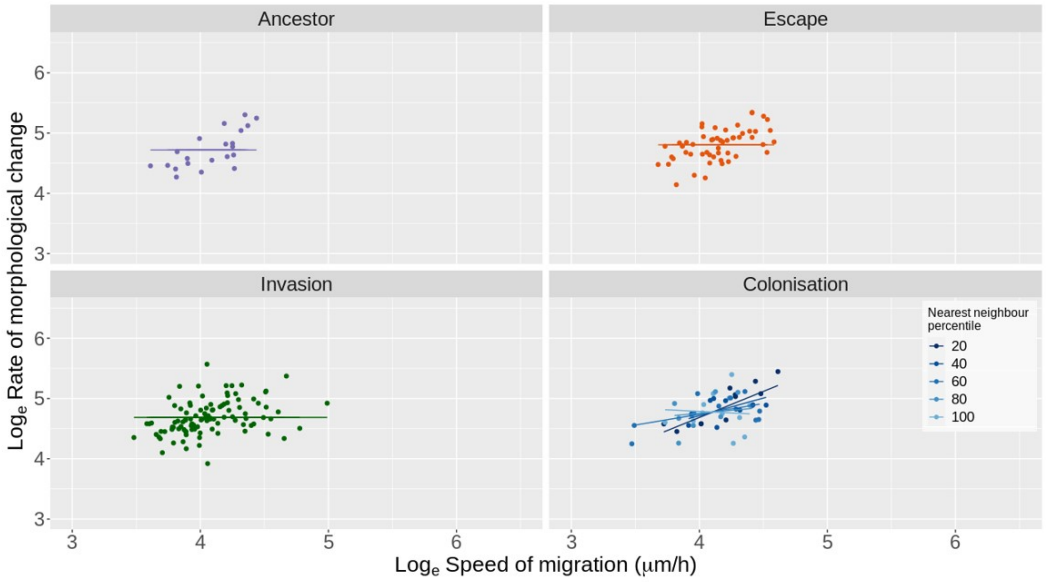


Figure B.3: The time invariant reduced model for each population after the removal of influential data points. The natural log-transformed rate of morphological change against the natural log-transformed speed of migration. In the colonisation populations the shaded lines indicate the natural log-transformed nearest neighbour percentile. The lighter the line the further away from a neighbouring cell. Influential data points, Cook's distance > (4 / N) where N is the sample size (Bollen and Jackman, 1985), have been removed to test whether a small subset of points influencing the result. After the removal of the influential points the intercept was still the only significant parameter in the ancestor, escape and invasion. Likewise, the speed of migration, distance to the nearest neighbouring cell and the interaction of the two was still significant in the colonisation populations.

Appendix C

Model selection

C.1 Chapter 4

Structure	Number of estimated parameters	Identification
Identity matrix	0	S1
A single global estimate	1	S2
Independent estimates for each video	33	S3
Independent estimates for each cell	244	S4
Correlated estimates for each video	66	S5
Correlated estimates for each cell within a given video	1265	S6

Table C.1: The observation variance structures compared during the dynamic factor model selection in Chapter 4. The 6 different observation variance structures compered during the dynamic factor model selection in Chapter 4. Note, the number of estimated parameters strictly related to the observation matrix structure and the not the model as a whole. The identification key relates to the observation variance structures in Tables C.2 - C.6.

Observation variance structure	Number of factors	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S5	4	-109605.92	221319.25	1040	0.00
S2	4	-109807.04	221588.14	975	267.89
S3	4	-109781.60	221602.89	1007	281.64
S4	4	-109624.39	221722.46	1218	400.21
S5	3	-110114.39	221840.87	798	517.63
S1	4	-110001.68	221975.37	974	651.13
S2	3	-110316.72	222113.01	733	787.76
S3	3	-110293.86	222132.51	765	806.26
S6	4	-108828.33	222263.56	2239	936.32
S4	3	-110159.30	222294.72	976	966.48
S5	2	-110632.67	222383.11	555	1053.86
S1	3	-110461.29	222400.10	732	1069.85
S2	2	-110835.79	222657.63	490	1326.39
S3	2	-110818.16	222687.19	522	1354.95
S6	3	-109341.59	222779.43	1997	1446.18
S1	2	-110935.69	222855.40	489	1521.15
S4	2	-110703.33	222886.22	733	1550.97
S6	2	-109861.26	223309.15	1754	1972.90
S5	1	-111354.05	223332.53	311	1995.28
S2	1	-111551.49	223596.49	246	2258.24
S3	1	-111540.14	223638.22	278	2298.97
S1	1	-111605.44	223702.39	245	2362.14
S4	1	-111456.38	223896.79	489	2555.55
S6	1	-110596.20	224270.50	1510	2928.25

Table C.2: The covariate free dynamic factor model selection results.
Displayed are the covariate free dynamic factor model selection results in Chapter 4. The observation variance structure relates to the identification column in Table C.1. The optimal covariate free model with the lowest AICc had 4 factors and correlated estimates for each video in the observation variance.

Observation variance structure	Number of factors	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S5	4	-110359.89	222827.19	1040	0.00
S2	4	-110430.97	222836.01	975	7.82
S3	4	-110415.27	222870.22	1007	41.03
S4	4	-110228.54	222930.76	1218	100.57
S1	4	-110568.01	223108.04	974	276.85
S5	3	-110798.73	223209.55	798	377.36
S2	3	-110872.47	223224.50	733	391.31
S3	3	-110858.25	223261.27	765	427.08
S4	3	-110701.35	223378.81	976	543.62
S1	3	-110972.14	223421.82	732	585.63
S6	4	-109482.55	223572.01	2239	734.82
S5	2	-111376.95	223871.66	555	1033.47
S2	2	-111449.76	223885.56	490	1046.37
S3	2	-111437.11	223925.08	522	1084.89
S6	3	-109945.35	223986.93	1997	1145.74
S1	2	-111510.91	224005.85	489	1163.66
S4	2	-111317.26	224114.08	733	1270.89
S6	2	-110545.75	224678.13	1754	1833.94
S5	1	-112196.15	225016.73	311	2171.54
S2	1	-112264.62	225022.75	246	2176.56
S1	1	-112287.71	225066.94	245	2219.75
S3	1	-112257.40	225072.75	278	2224.56
S4	1	-112220.75	225425.52	489	2576.33
S6	1	-111392.45	225863.00	1510	3012.81

Table C.3: The speed of migration dynamic factor model selection results.
Displayed are the speed of migration dynamic factor model selection results used in Chapter 4 to impute the missing speed of migration values. The observation variance structure relates to the identification column in Table C.1. The optimal model used for imputation had 4 factors and correlated estimates for each video in the observation variance. The imputed speed of migration values were then used in Chapter 4 and Chapter 5.

Observation variance structure	Number of factors	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S6	4	-57082.76	118772.53	2239	0.00
S6	3	-63095.08	130286.48	1997	11512.95
S4	4	-68491.33	139456.35	1218	20681.82
S6	2	-70133.26	143853.21	1754	25077.68
S5	4	-74402.09	150911.60	1040	32135.07
S3	4	-74815.14	151669.98	1007	32892.45
S4	3	-75622.33	153220.79	976	34442.26
S2	4	-76256.91	154487.91	975	35708.38
S6	1	-77626.35	158330.84	1510	39550.31
S5	3	-81181.81	163975.73	798	45194.20
S3	3	-81553.41	164651.60	765	45869.07
S2	3	-82763.81	167007.20	733	48223.67
S4	2	-84258.38	169996.34	733	51211.80
S5	2	-88713.66	178545.09	555	59759.56
S3	2	-89139.94	179330.75	522	60544.22
S2	2	-89905.88	180797.82	490	62010.29
S1	4	-89437.53	180847.10	974	62058.57
S1	3	-92185.48	185848.50	732	67058.97
S4	1	-94685.34	190354.71	489	71564.18
S1	2	-95747.25	192478.53	489	73687.00
S5	1	-97887.61	196399.65	311	77607.12
S3	1	-98417.58	197393.11	278	78599.57
S2	1	-98837.81	198169.14	246	79374.61
S1	1	-101250.27	202992.05	245	84196.52

Table C.4: The nearest neighbour dynamic factor model selection results.
Displayed are the nearest neighbour dynamic factor model selection results used in Chapter 4 to impute the missing nearest neighbour distances. The observation variance structure relates to the identification column in Table C.1. The optimal model used for imputation had 4 factors and correlated estimates for each cell within a given video. The imputed nearest neighbour distances were then used in Chapter 4 and 5.

Observation variance structure	Covariate combination	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S4	Full	-100758.67	204015.76	1230	0.00
S4	Both	-100849.90	204189.97	1226	173.21
S4	Speed	-100889.71	204261.34	1222	243.58
S5	Full	-101126.98	204386.00	1052	367.24
S5	Both	-101202.11	204528.05	1048	508.29
S6	Full	-99970.33	204572.98	2251	552.22
S5	Speed	-101240.89	204597.40	1044	575.64
S6	Both	-100053.34	204730.53	2247	707.77
S3	Full	-101346.50	204757.30	1019	733.53
S6	Speed	-100091.28	204797.93	2243	773.17
S2	Full	-101446.96	204892.59	987	866.83
S3	Both	-101420.73	204897.56	1015	870.80
S3	Speed	-101459.55	204966.99	1011	939.23
S2	Both	-101519.72	205029.91	983	1001.14
S2	Speed	-101557.76	205097.79	979	1068.03
S1	Full	-103169.03	208334.68	986	4303.92
S1	Both	-103222.77	208433.95	982	4402.19

Table C.5: The dynamic factor model selection results with covariates estimated at a population level. Displayed are the dynamic factor model selection results in Chapter 4 with covariate effects estimated at a population level. The observation variance structure relates to the identification column in Table C.1. The covariate combination relates to either: the speed of migration only (speed), the speed of migration and the nearest neighbour distance (both), or the full model of covariate effects (full). All of the models were estimated with 4 factors. The optimal dynamic factor model with covariate effects estimated at a population level had a full model of covariates with an independent estimates for each cell in the observation matrix.

Observation variance structure	Covariate combination	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S4	Full	-99306.94	202611.30	1950	0.00
S4	Both	-99734.70	202955.74	1706	343.44
S5	Full	-99681.37	202987.00	1772	373.70
S4	Speed	-100064.12	203106.68	1462	492.37
S6	Full	-98517.35	203205.80	2971	590.50
S5	Both	-100086.65	203288.80	1528	672.50
S3	Full	-99905.13	203365.54	1739	748.24
S5	Speed	-100409.62	203429.13	1284	810.83
S6	Both	-98937.52	203521.45	2727	902.15
S2	Full	-100019.49	203527.41	1707	907.11
S6	Speed	-99271.22	203667.48	2483	1046.18
S3	Both	-100310.92	203668.78	1495	1046.48
S3	Speed	-100629.59	203800.94	1251	1177.64
S2	Both	-100425.26	203831.03	1463	1206.73
S2	Speed	-100737.54	203950.81	1219	1325.51
S1	Full	-102137.00	207760.34	1706	5134.04
S1	Speed	-102652.21	207778.08	1218	5150.78
S1	Both	-102423.86	207826.15	1462	5197.85
S5	Neighbour	-109184.49	220978.88	1284	18349.58
S2	Neighbour	-109393.39	221262.51	1219	18632.21
S3	Neighbour	-109364.41	221270.58	1251	18639.28
S4	Neighbour	-109197.83	221374.10	1462	18741.80
S1	Neighbour	-109629.11	221731.89	1218	19098.59
S6	Neighbour	-108404.36	221933.74	2483	19299.44

Table C.6: The dynamic factor model selection results with covariates estimated at a single cell level. Displayed are the dynamic factor model selection results in Chapter 4 with covariate effects estimated at a single cell level. The observation variance structure relates to the identification column in Table C.1. The covariate combination relates to either: the speed of migration only (speed), the speed of migration and the nearest neighbour distance (both), or the full model of covariate effects (full). All of the models were estimated with 4 factors. The optimal dynamic factor model with covariate effects estimated at a single cell level had a full model of covariates with an independent estimate for each cell in the observation matrix.

C.2 Chapter 5

Structure	Number of estimated parameters	Identification
Identity matrix	0	S1
A single global estimate	1	S2
Independent estimates for each population	3	S3
Independent estimates for each video	8	S4
Independent estimates for each cell	69	S5
Correlated estimates for each video	16	S6
Correlated estimates for each cell within a given video	432	S7

Table C.7: The state matrix and state variance structures compared during the blind model selection in Chapter 5. The 7 different parameter structures compared in the state matrix and state variance blind model selection in Chapter 5. Note, the number of estimated parameters strictly relate to the parameter structure and not the model as a whole. The identification key relates to the state matrix and state variance structures in Tables C.8 and C.9.

State matrix structure	State variance structure	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S4	S4	-30872.94	62023.58	138	0.00
S4	S5	-31000.12	62166.86	83	142.27
S2	S5	-31072.04	62174.09	15	148.51
S31	S5	-31070.18	62174.39	17	147.81
S4	S2	-31019.26	62178.95	70	151.37
S4	S31	-31019.16	62182.79	72	154.21
S4	S3	-31018.24	62188.99	76	159.41
S31	S2	-31091.33	62190.67	4	160.09
S2	S31	-31093.09	62194.19	4	162.60
S31	S31	-31091.28	62194.57	6	161.99
S2	S3	-31089.31	62194.63	8	161.05
S31	S3	-31087.43	62194.86	10	160.28
S2	S2	-31097.77	62199.55	2	163.96
S4	S6	-30635.14	62274.01	491	237.43
S31	S4	-31068.36	62281.19	72	243.60
S2	S4	-31070.74	62281.92	70	243.34
S31	S6	-30801.16	62468.57	425	428.99
S2	S6	-30804.18	62470.46	423	429.87
S31	S1	-33217.79	66441.58	3	4399.99
S4	S1	-33153.48	66445.40	69	4402.81
S2	S1	-33222.96	66447.92	1	4404.33

Table C.8: The covariate free blind model selection results.
Displayed are the covariate free blind model selection results in Chapter 5. The state matrix and state variance structure relate to the identification column in Table C.7. All of the models were estimated with a fixed observation variance as detailed in Section 5.3.2. The optimal covariate free blind model with the lowest AICc had independent estimates in the state matrix and the state variance.

State variance structure	Covariate combination	Log likelihood	AICc	Number of estimated parameters	ΔAICc
S4	Full	-28686.27	58073.22	345	0.00
S5	Full	-28755.12	58097.76	290	23.54
S2	Full	-28786.57	58134.00	277	58.78
S31	Full	-28785.27	58135.50	279	59.28
S3	Full	-28782.81	58138.79	283	61.56
S4	Both	-28800.26	58159.34	276	81.12
S5	Both	-28872.42	58191.20	221	111.98
S2	Both	-28901.60	58223.06	208	142.84
S31	Both	-28901.02	58225.98	210	144.75
S3	Both	-28898.76	58229.61	214	147.38
S4	Speed	-28983.82	58385.47	207	302.25
S5	Speed	-29040.53	58387.13	152	302.90
S6	Full	-28506.88	58454.09	698	368.86
S2	Speed	-29088.15	58456.01	139	369.79
S31	Speed	-29086.96	58457.69	141	370.47
S3	Speed	-29085.63	58463.14	145	374.91
S6	Both	-28611.12	58516.13	629	426.91
S6	Speed	-28757.47	58663.30	560	573.08
S4	Neighbour	-30748.48	61914.79	207	3823.57
S5	Neighbour	-30858.20	62022.46	152	3930.24
S2	Neighbour	-30873.34	62026.40	139	3933.17
S31	Neighbour	-30872.89	62029.55	141	3935.32
S3	Neighbour	-30871.30	62034.48	145	3939.25
S6	Neighbour	-30503.06	62154.47	560	4058.25
S1	Both	-31839.34	64096.50	207	5999.27
S1	Full	-31789.39	64137.59	276	6039.37
S1	Speed	-31930.70	64139.10	138	6039.88
S1	Neighbour	-33011.10	66299.89	138	8199.67

Table C.9: The blind model covariate selection results.
Displayed are the blind model covariate selection results in Chapter 5. The state variance structure relates to the identification column in Table C.7. The covariate combination relates to either: the speed of migration only (speed), the speed of migration and the nearest neighbour distance (both), or the full model of covariate effect (full). All of the models were estimated with independent estimates for each cell in the state matrix. Likewise, all of the models were estimated with a fixed observation variance as detailed in Section 5.3.2. The optimal blind model had a full model of covariates with independent estimates for each cell in the state variance.