PHD STUDENTSHIP PROJECT PROPOSAL:

PROJECT DETAILS	
Project Title:	Soft Tissue Sarcoma Diagnosis and Classfication from Histopathological Images using Deep Learning
Short Project Title:	Using Deep Learning to Diagnose Sarcoma
SUPERVISORY TEAM	
Primary Supervisor:	Chris Bakal
Additional members of the supervisory team:	Khin Thway, Janet Shipley
DIVISIONAL AFFILIATION	
Primary Division:	Cancer Biology
Primary Team:	Dynamical Cell Systems
Other Division (if applicable):	Division of Molecular Pathology
Other Team (if applicable):	Sarcoma Unit, The Royal Marsden NHS Foundation Trust

PROJECT PROPOSAL

BACKGROUND TO THE PROJECT

Soft Tissue Sarcomas (STS) are a highly heterogeneous group of cancers that originate from fat, muscle, nerves, fibrous tissues, or blood vessels; with variable levels of maturity, and accounting for >50 basic known histologies. Each subtype has a spectrum of morphological appearances, and there is much histologic and immunophenotypic overlap between STS tumours and other neoplasms such as carcinomas, lymphomas and melanomas. Additionally, for every STS observed, 100 benign tumours present; and benign and malignant neoplasms can share highly similar morphologic patterns (Fletcher, 2006). ***The diagnosis of STS is therefore often regarded as the most complex of all cancer diagnoses.***

All STS cases are referred to highly specialized pathology teams. Yet, even with expert histological analysis, around 15% of sarcomas remain unclassifiable. Failure to accurately diagnose STS leads to 'trial and error' treatment which is non-tailored and non-specific (Ducimetiere et al., 2011). Although approaches to diagnose and classify STS based on genomic alterations, especially translocations, are rapidly evolving – these are still in relatively early stages compared to their use in other cancer types (Mertens and Tayebwa, 2014, Thway et al., 2010, Thway et al., 2015). ***As categorization is difficult even for highly***

expert STS pathologists, there is an urgent unmet need for digital pathology tools by which these tumours can be diagnosed and classified, leading to more accurate patient stratification for treatment.

Deep learning approaches have shown super-human performance in a range of image classification challenges, and are set to revolutionize biomedical research, clinical practice, and the healthcare industry (Litjens et al., 2017). Deep learning methods have already been used to successfully identify and distinguish normal from tumour breast tissue (Xu et al., 2017), as well as detecting invasive versus non invasive breast tumours (Cruz-Roa et al., 2017). Yet, to date deep learning approaches have only been applied to routine analysis of common cancers. ***In this project, a computational graduate student will develop deep learning methods that can be applied to the diagnosis of STS.***

PROJECT AIMS

- **Aim 1. Generate a digital repository of STS tumour sections images at the Institute of Cancer Research (Assembly: Months 1-12, Maintenance: Months 12-48).**
- **Aim 2: Developing deep-learning methods to diagnosis STS (Months 12-36).**
- **Aim 3: Developing methods to predict genomic alterations in STS (Months 36-48).**

RESEARCH PROPOSAL

Aim 1. Generate a digital repository of STS tumour sections images at the Institute of Cancer Research (Assembly: Months 1-12, Maintenance: Months 12-48).

Initially the student will spend significant time generating a digital image repository of STS samples. This will involve imaging of each sample, and then creating a searchable database where images can be found using different descriptors (sub-type, grade, positivity for biomarkers such as Glypican-3, pathologist notes). This database will not contain patient data.

The student first make use of a previously assembled set of samples (Thway et al., 2011) where cores have been taken from 5 different classes. While putting together this database is expected to be time consuming, it will be necessary to perform deep learning studies. ***Moreover, this database will also be an invaluable resource for sarcoma researchers in the UK and around the world who wish to query the data for additional studies.***

This dataset has already undergone quality control and subsequent annotation by expert pathologists including Dr. Thway. Moreover, Dr. Thway will supervise the student in the assembly of this database, and train the student in basic aspects of pathology workflows in order to ensure no errors (i.e. mislabeling, introduction of poor quality images) are made during assembly of the dataset.

In the second phase of the project, the student will add additional samples to the dataset. These are patient samples where cores have been taken, and where an NGS panel is being used to characterize genomic alterations. These samples are being generated as part of a project led by David Gonzalez de Castro (Queen's University, Belfast), and a Sarcoma UK Pilot award. Additionally, we have established a collaboration with the STS project of International Cancer Genome Consortium, who will provide additional images of tumours and their associated sequence data.

Aim 2: Developing deep-learning methods to diagnosis STS (Months 12-36).

Despite the size and diversity of this dataset the student will assemble in Aim 1, it remains small for many conventional deep learning approaches. To address this challenge the student will initially build on advances in the field of deep "one-shot" learning (Lake et al., 2015). The Bakal laboratory has already used one-shot deep learning methods to predict breast cancer grade with similar accuracy levels to pathologists (Sam Cooper et al., unpublished).

However, should the student not be able to achieve successful classification using one-shot deep learning methods, we may consider using more classical, though laborious, methods of image analysis and machine learning.

Aim 3: Developing methods to predict genomic alterations in STS (Months 36-48).

Whilst the goal of Aim 2 is to predict STS subtypes and grades that were based on visual examination by expert pathologists, in Aim 3, the student will develop deep learning methods to predict sub-types determined by NGS sequencing approaches. Meaning that the student will aim to predict genotypes from phenotypes. Recently, the Bakal laboratory has developed conceptually similar methodology to predict gene expression from breast tumour cell morphology (Sailem and Bakal, 2017).

Here the student will leverage data generated as part of a Sarcoma Research UK and CRUK funded project led by led by David Gonzalez de Castro (Queen's University, Belfast), which has collected 100 sarcoma samples (archival FFPE blocks), where FISH and/or RT-PCR has identified the presence of a fusion, mutation, or *MDM2* amplification, and are now being analysed for genomic alterations in a panel of genes by NGS. In addition, we have established a collaboration with the Soft Tissue Cancer Project of the International Cancer Genome Consortium (ICGC), led by Frédéric Chibon, who will also provide images of sarcoma tumours that are also being profiled using whole-genome sequencing. The ICGC has compiled a sample set of almost 1000 tumours, and the first phase of the project involves genome sequencing of ~150.

As the size of this dataset is even smaller than that used in Aim 2, deep learning based methods may not be appropriate to use here. However, to ensure a successful outcome to the study, the student will employ more conventional machine learning approaches in order to classify different sub-types such as classical Neural Network, Support Vector Machine (SVM), or Random Forest based methods. The Bakal laboratory has extensive expertise in the use of such techniques. For example, we have used classical Neural Network based methods to classify cell morphology (Bakal et al., 2007), or SVMs to describe phenotypic heterogeneity (Yin et al., 2013).

EXPLAIN BRIEFLY HOW THIS PROJECT WILL PROVIDE A TRANSLATIONAL TRAINING EXPERIENCE FOR THE STUDENT

By working on a project to bring new image analysis methods into the clinic which will directly impact the well-being of patients, this project exemplifies a translational training experience. The student will be co-supervised by Dr. Chris Bakal, Professor Janet Shipley – team leaders at Institute of Cancer Research (ICR), and by Dr. Khin Thway a Consultant Pathologist at the Royal Marsden Hospital. Working with Dr. Bakal, the student will develop skills in image analysis and computational biology; which will be integrated with extensive training provided by Professor Shipley and Dr. Thway in the pathology and diagnosis of sarcomas that are being treated at the Royal Marsden**. Thus, the student will be in the unique position of being able to develop new computational tools to affect sarcoma diagnosis in a clinical setting.** The student will also have the unique opportunity to interact with the laboratory of Professor David Gonzalez de Castro at Queen's University Belfast, and the STS Project of the International Cancer Genome Consortium based in Toulouse, France and managed by Dr. Frédéric Chibon.

LITERATURE REFERENCES

BAKAL, C., AACH, J., CHURCH, G. & PERRIMON, N. 2007. Quantitative morphological signatures define local signaling networks regulating cell morphology. *Science,* 316**,** 1753-6.

CRUZ-ROA, A., GILMORE, H., BASAVANHALLY, A., FELDMAN, M., GANESAN, S., SHIH, N. N. C., TOMASZEWSKI, J., GONZALEZ, F. A. & MADABHUSHI, A. 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: A Deep Learning approach for quantifying tumor extent. *Sci Rep,* 7**,** 46450.

DUCIMETIERE, F., LURKIN, A., RANCHERE-VINCE, D., DECOUVELAERE, A. V., PEOC'H, M., ISTIER, L., CHALABREYSSE, P., MULLER, C., ALBERTI, L., BRINGUIER, P. P., SCOAZEC, J. Y., SCHOTT, A. M., BERGERON, C., CELLIER, D., BLAY, J. Y. & RAY-COQUARD, I. 2011. Incidence of sarcoma histotypes and molecular subtypes in a prospective epidemiological study with central pathology review and molecular testing. *PLoS One,* 6**,** e20294.

FLETCHER, C. D. 2006. The evolving classification of soft tissue tumours: an update based on the new WHO classification. *Histopathology,* 48**,** 3-12.
LAKE, B. M., SALAKHUTDINOV, R. & TENENBAUM, J. B. 2015. Human-level concept learning through probabilistic program induction. *Science,* 350**,** 1332-8.
LITJENS, G., KOOI, T., BEJNORDI, B. E., SETIO, A. A. A., CIOMPI, F., GHAFOORIAN, M., VAN DER LAAK, J., VAN GINNEKEN, B. & SANCHEZ, C. I. 2017. A survey on deep learning in medical image analysis. *Med Image Anal,* 42**,** 60-88.
MERTENS, F. & TAYEBWA, J. 2014. Evolving techniques for gene fusion detection in soft tissue tumours. *Histopathology,* 64**,** 151-62.
SAILEM, H. Z. & BAKAL, C. 2017. Identification of clinically predictive metagenes that encode components of a network coupling cell shape to transcription by image-omics. *Genome Res,* 27**,** 196-207.
THWAY, K., ROCKCLIFFE, S., GONZALEZ, D., SWANSBURY, J., MIN, T., THOMPSON, L. & FISHER, C. 2010. Utility of sarcoma-specific fusion gene analysis in paraffin-embedded material for routine diagnosis at a specialist centre. *J Clin Pathol,* 63**,** 508-12.
THWAY, K., SELFE, J., MISSIAGLIA, E., FISHER, C. & SHIPLEY, J. 2011. Glypican-3 is expressed in rhabdomyosarcomas but not adult spindle cell and pleomorphic sarcomas. *J Clin Pathol,* 64**,** 587-91.
THWAY, K., WANG, J., WREN, D., DAINTON, M., GONZALEZ, D., SWANSBURY, J. & FISHER, C. 2015. The comparative utility of fluorescence in situ hybridization and reverse transcription-polymerase chain reaction in the diagnosis of alveolar rhabdomyosarcoma. *Virchows Arch,* 467**,** 217-24.
XU, Y., JIA, Z., WANG, L. B., AI, Y., ZHANG, F., LAI, M. & CHANG, E. I. 2017. Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features. *BMC Bioinformatics,* 18**,** 281.
YIN, Z., SADOK, A., SAILEM, H., MCCARTHY, A., XIA, X., LI, F., GARCIA, M. A., EVANS, L., BARR, A. R., PERRIMON, N., MARSHALL, C. J., WONG, S. T. & BAKAL, C. 2013. A screen for morphological complexity identifies regulators of switch-like transitions between discrete cell shapes. *Nat Cell Biol,* 15**,** 860-71.

CANDIDATE PROFILE

Note: the ICR's standard minimum entry requirement is a relevant undergraduate Honours degree (First or 2:1)

Pre-requisite qualifications of applicants: e.g. BSc or equivalent in specific subject area(s)	BSc in Computer Science, Engineering, Mathematics, Physics, Statistics, or Bioinformatics. Candidates from Biology backgrounds with extensive experience in coding will also be considered.
Intended learning outcomes: Please provide a bullet point list (maximum of seven) of the knowledge and skills you expect the student to have attained on completion of the project.	• Extensive skills in using deep learning and other artificial intelligence methods to analysis tumour sections. • Experience in digital pathology. • Expertise in visual diagnosis of sarcoma. • Knowledge in sarcoma biology and treatment. • Skills in integration of imaging, NGS, and clinical data.