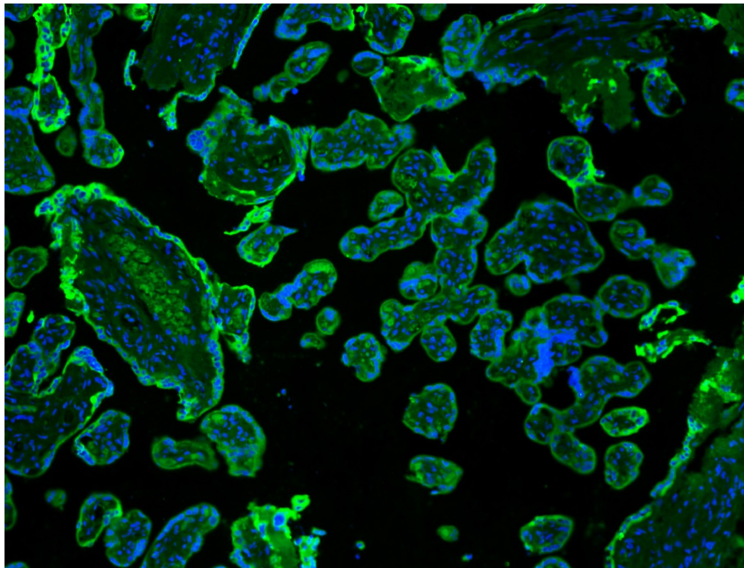
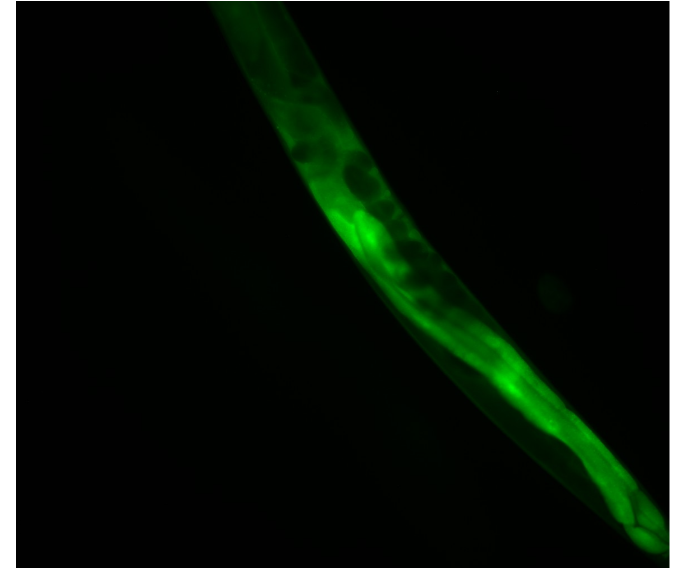
Image Data Analysis in H. Sapiens and C. Elegans



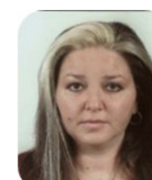


Dr. Alexander K. Seewald

SEEWALD SOLUTIONS

(Our) Relevant Collaborations for Image Data Analysis

2009-2012
Med.Univ.Vienna, Austria
Tissue Gnostics GmbH
Funded by FFG Bridge



2007-2010
Univ. Colorado@Boulder, USA
Self-funded

2011-
IMP Vienna

© 2012

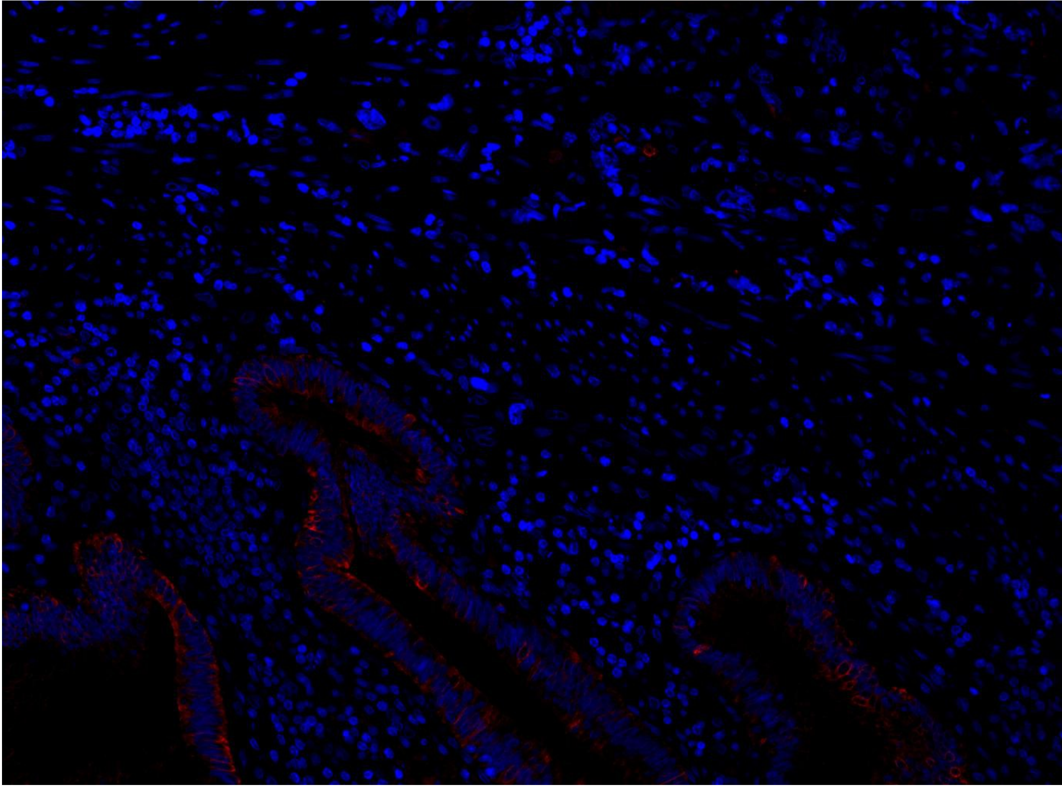
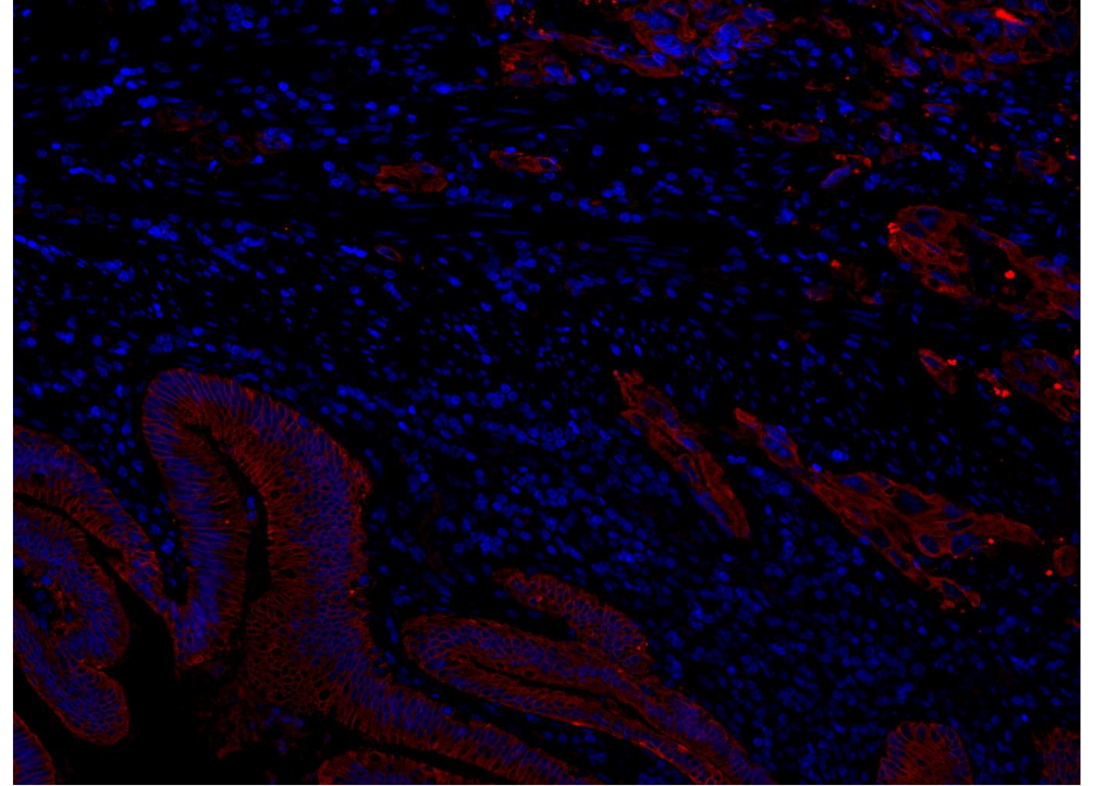
Science2Business 2011 Award (3rd Place)

Video 1 - How Tissue Samples Are Created

© 2012

SEEWALD SOLUTIONS
alex@seewald.at
www.seewald.at

Staining tissue is a complex process!

- Tissue characteristics (thickness, density, age, photobleaching, ...)
- Quality of antibodies (too old, too new, wrong kind, cross-reactions, ..)
- Crosstalk between multiple stainings, Protocols, MTA characteristics...

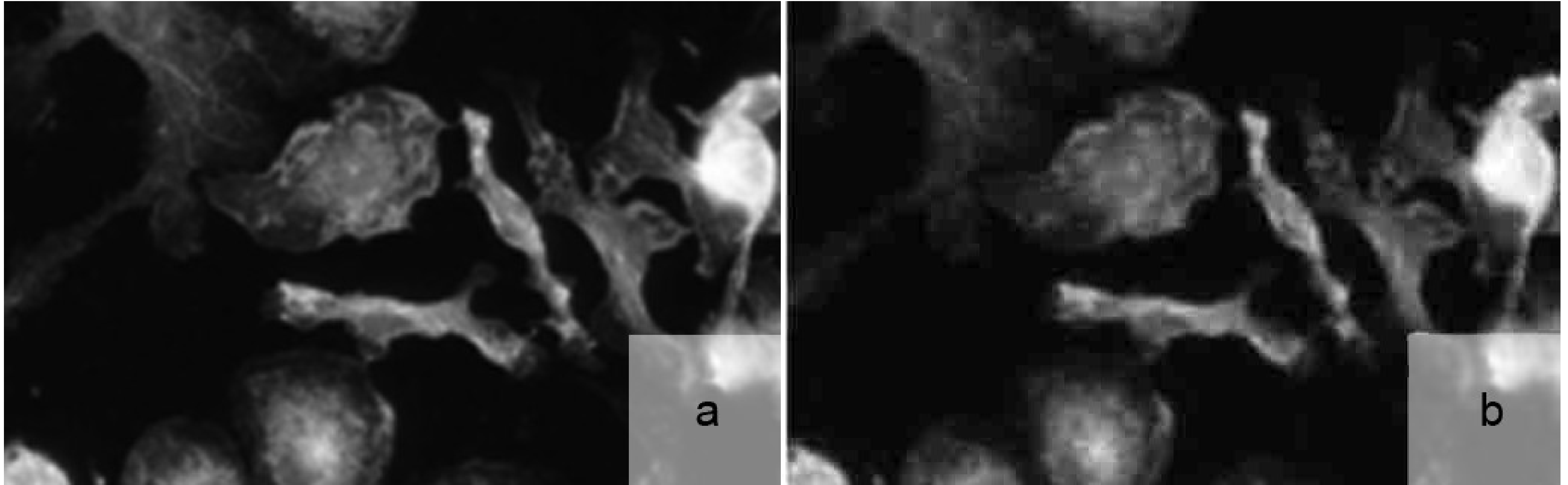
© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Image Formats

Use only lossless formats (TIFF, PNG)!

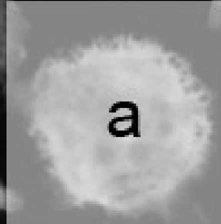


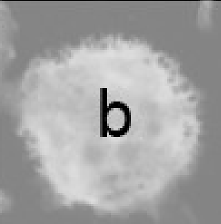
Especially JPEG is to be avoided! Removes information not visible to human eye which is still useful to image processing, introduces artefacts... **Also: No normalization of input images!**

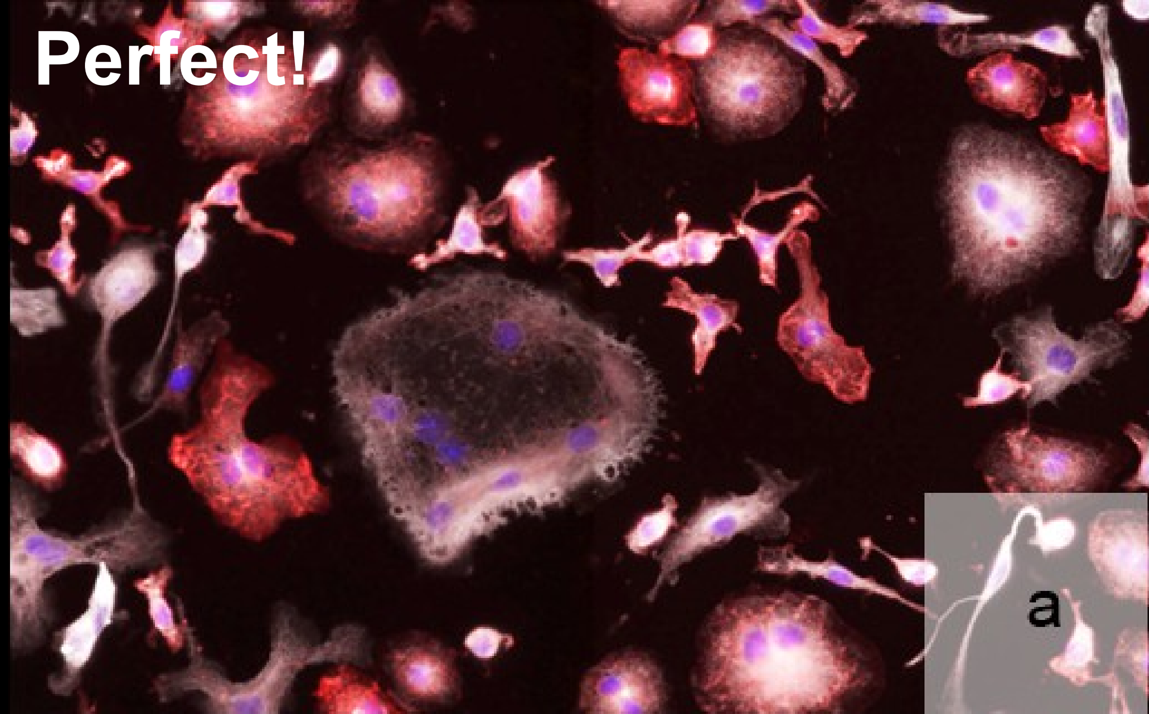
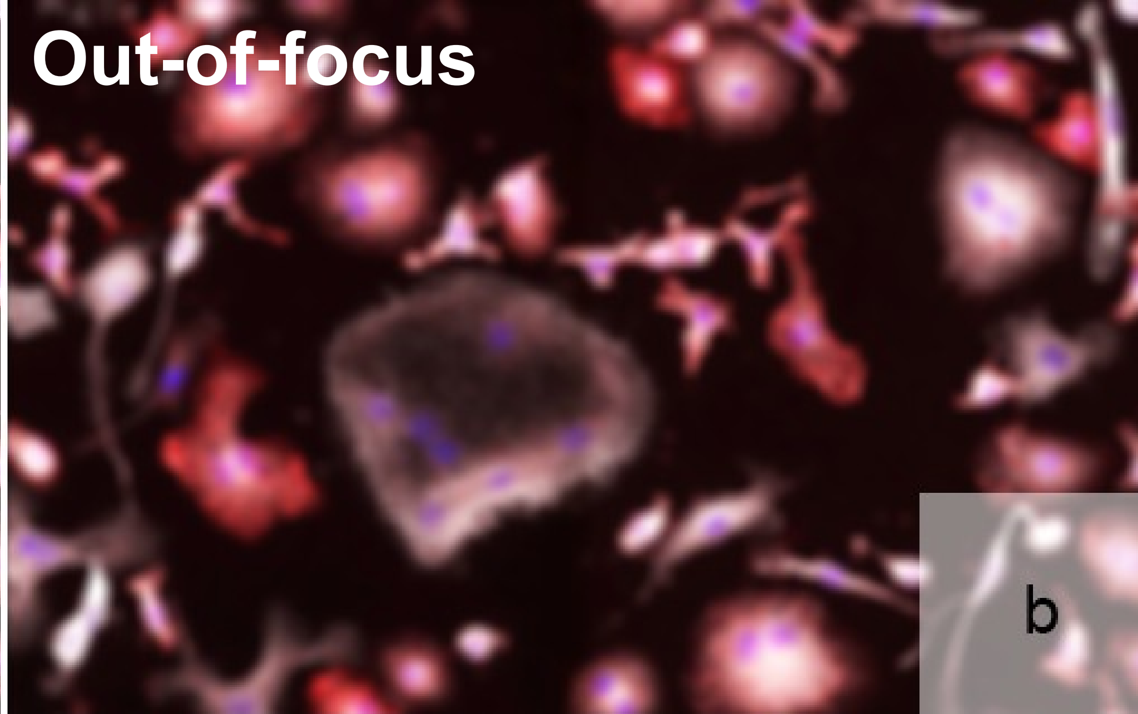
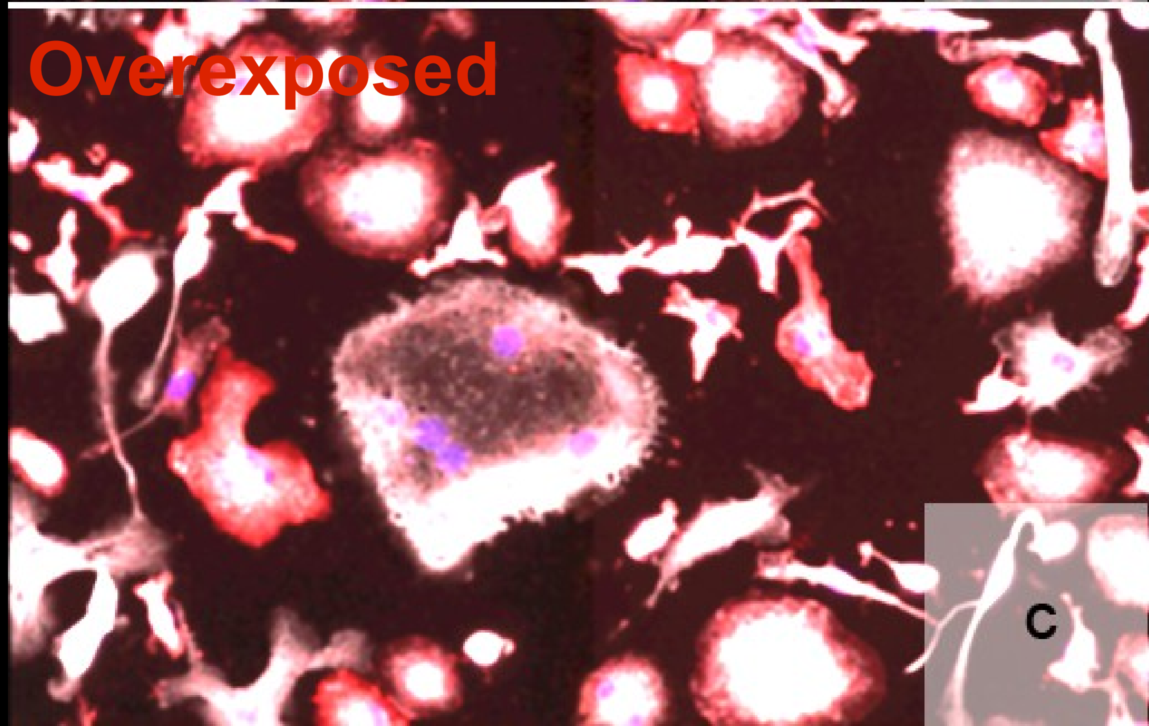
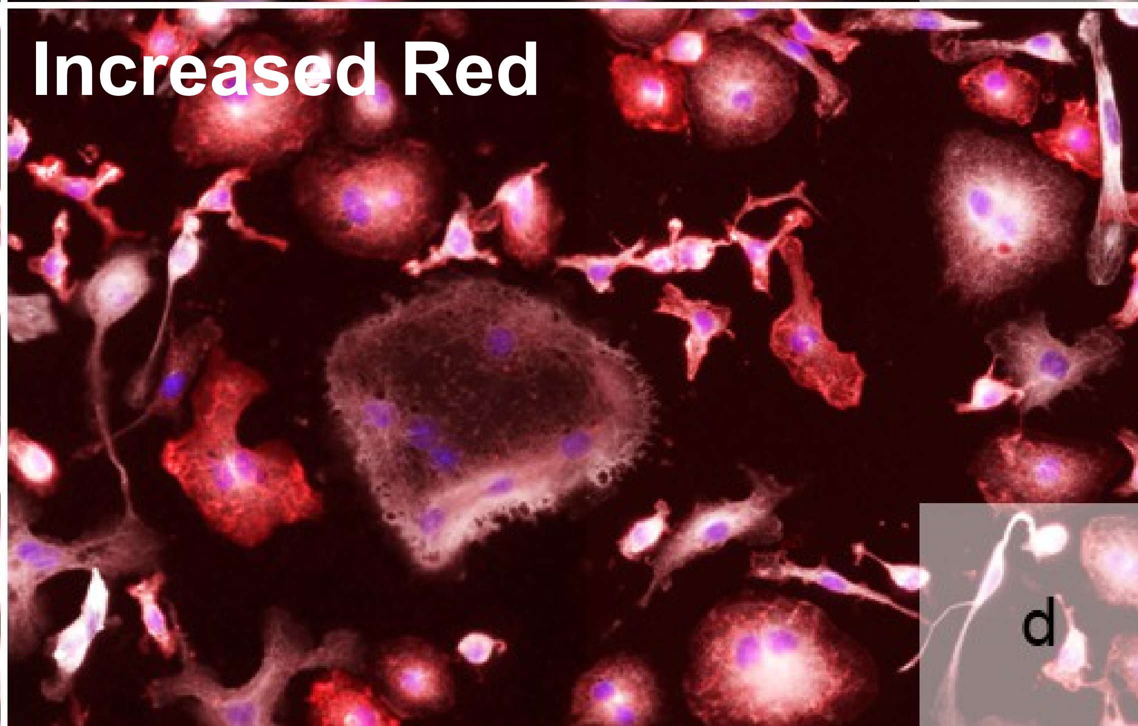
© 2012

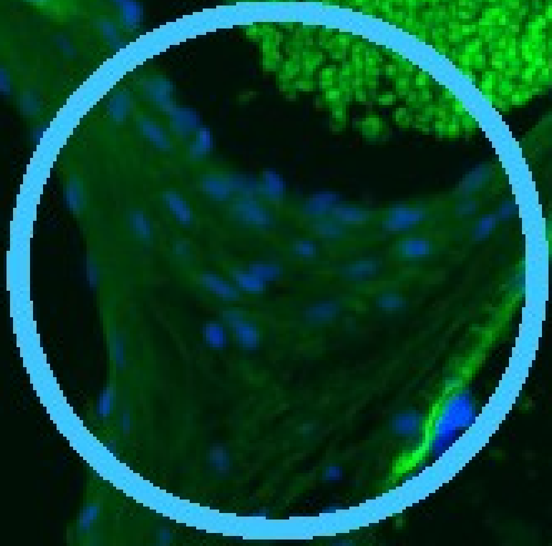
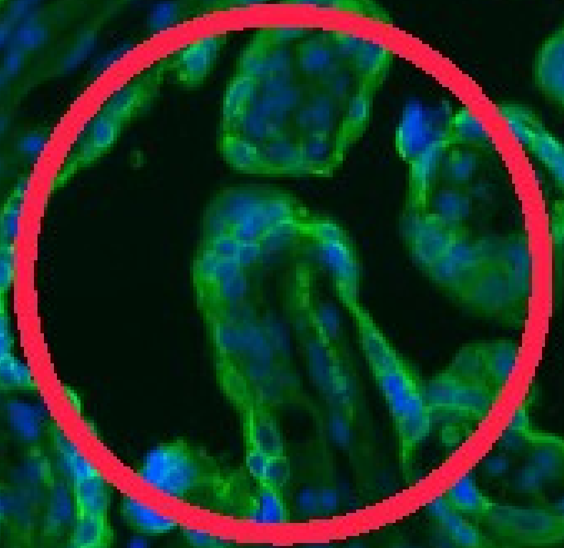
alex@seewald.at
www.seewald.at

Uneven Illumination






Perfect!
a

Out-of-focus
b

Overexposed
c

Increased Red
d

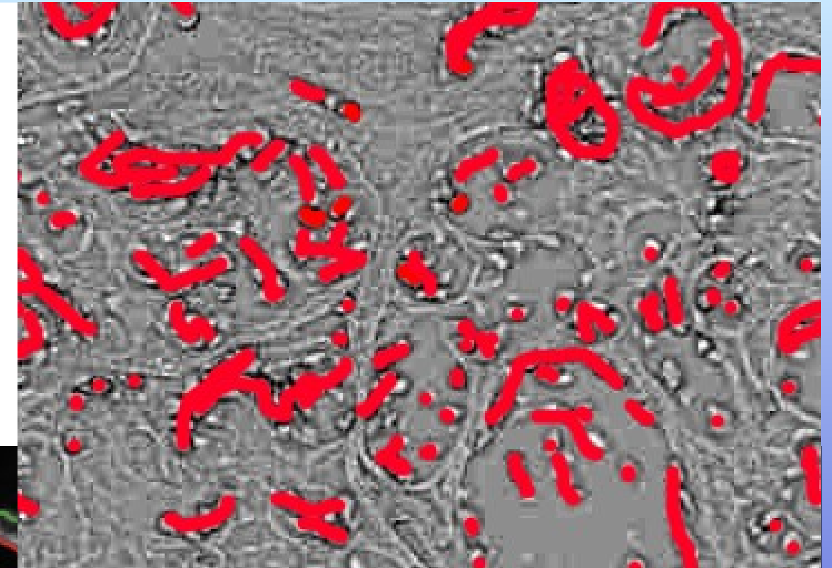



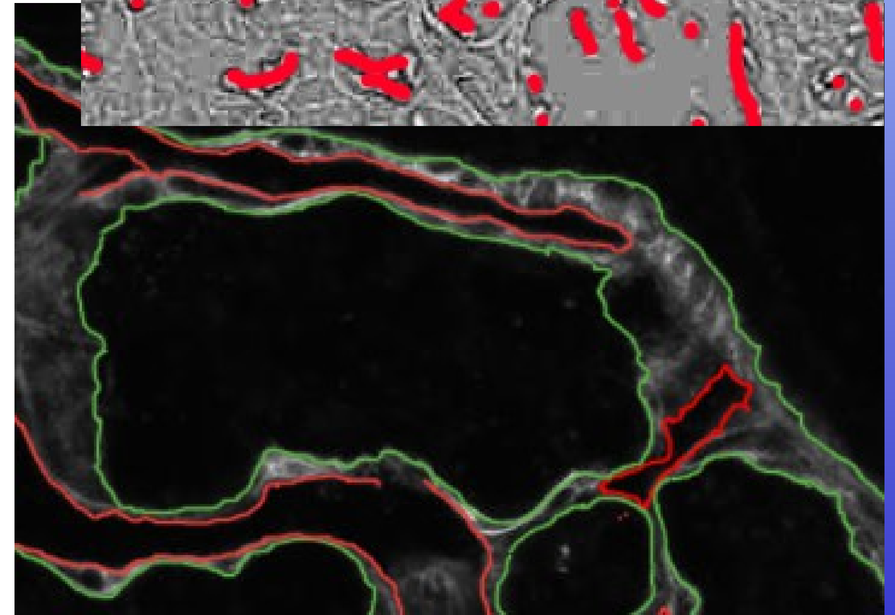
Slide-based microscopy records many images by automatically moving the tissue sample vs. the microscope. Tiling images together may not be sufficient.
Stitching analyzes an over-lap between adjacent images to deter-mine a more precise combination. **We had to develop our own algorithm.**

Ground-Truth Markup (1)

Image processing systems should be designed by biological experts – not by computer scientists with some biological background!

One way to do this is by convincing biological experts to create ground-truth markup.





Creating Markups

- Use any painting program (Photoshop,GIMP,...)
- Open original image, add new layer for markup
- Save in multi-layered format (TIFF, XCF)
- Digitizer tablets improve speed & accuracy
- Use „average“ images, not the best ones!

© 2012

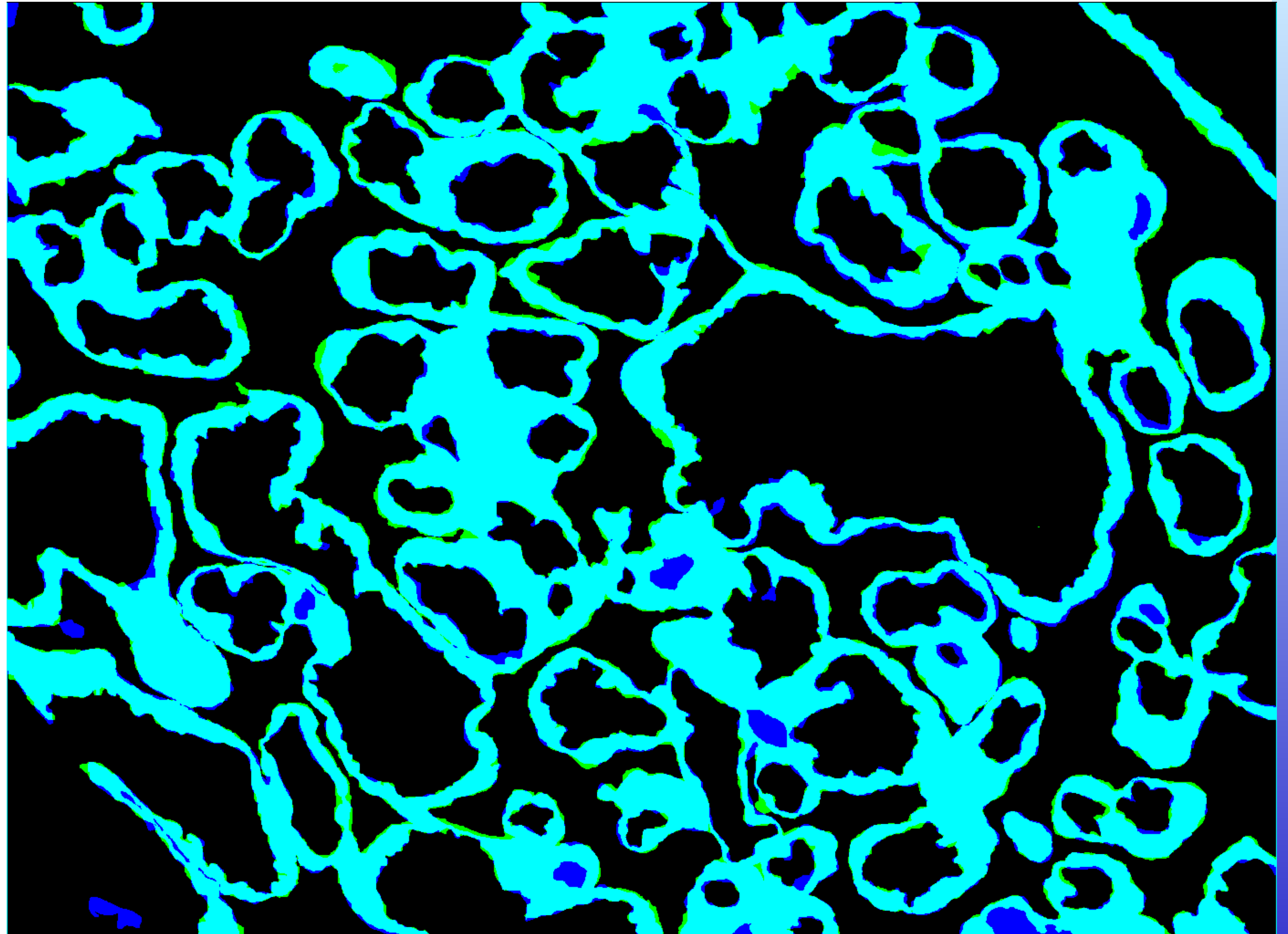
alex@seewald.at
www.seewald.at

Ground-Truth Markup (2)

Ground-Truth Markup is very useful in many ways

- Compute inter-expert variance
  e.g. right:Syncytiotrophoblast
- Determine hardness of task
- Validate & finetune IA systems
- Optimize parameters rather than using ad-hoc values
- Train IA system using Machine Learning (*Ery-Removal* - later)

...

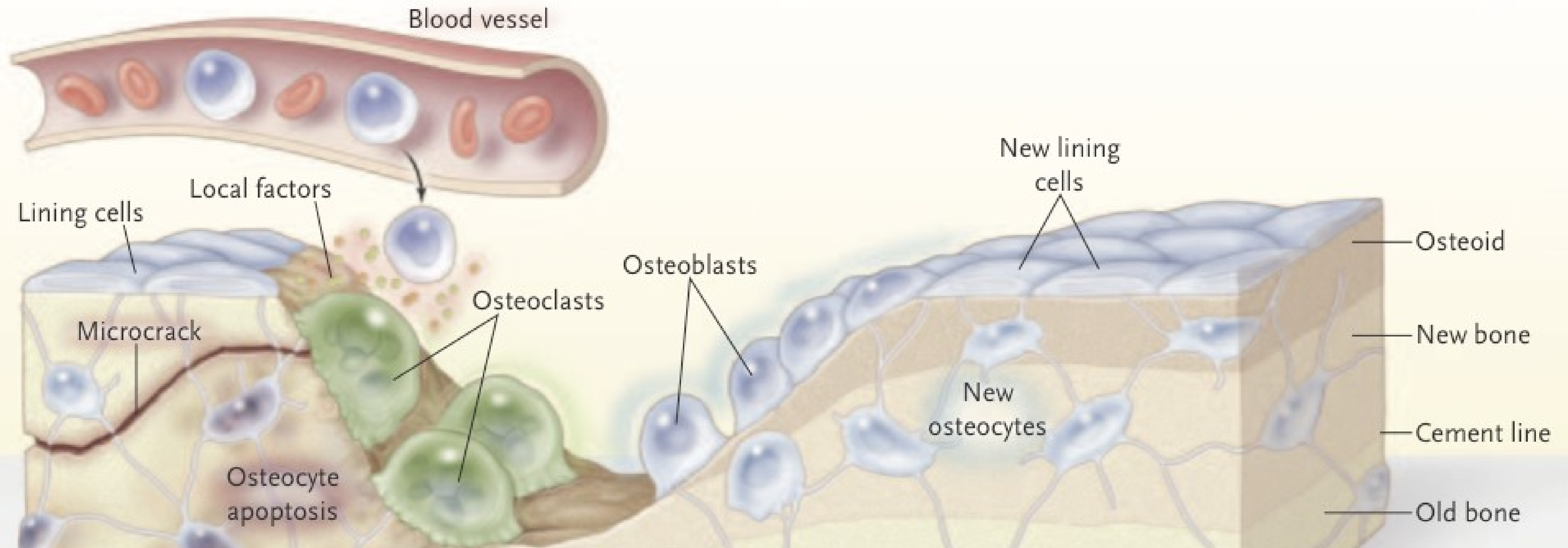


© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Osteoclast Characterization (1)



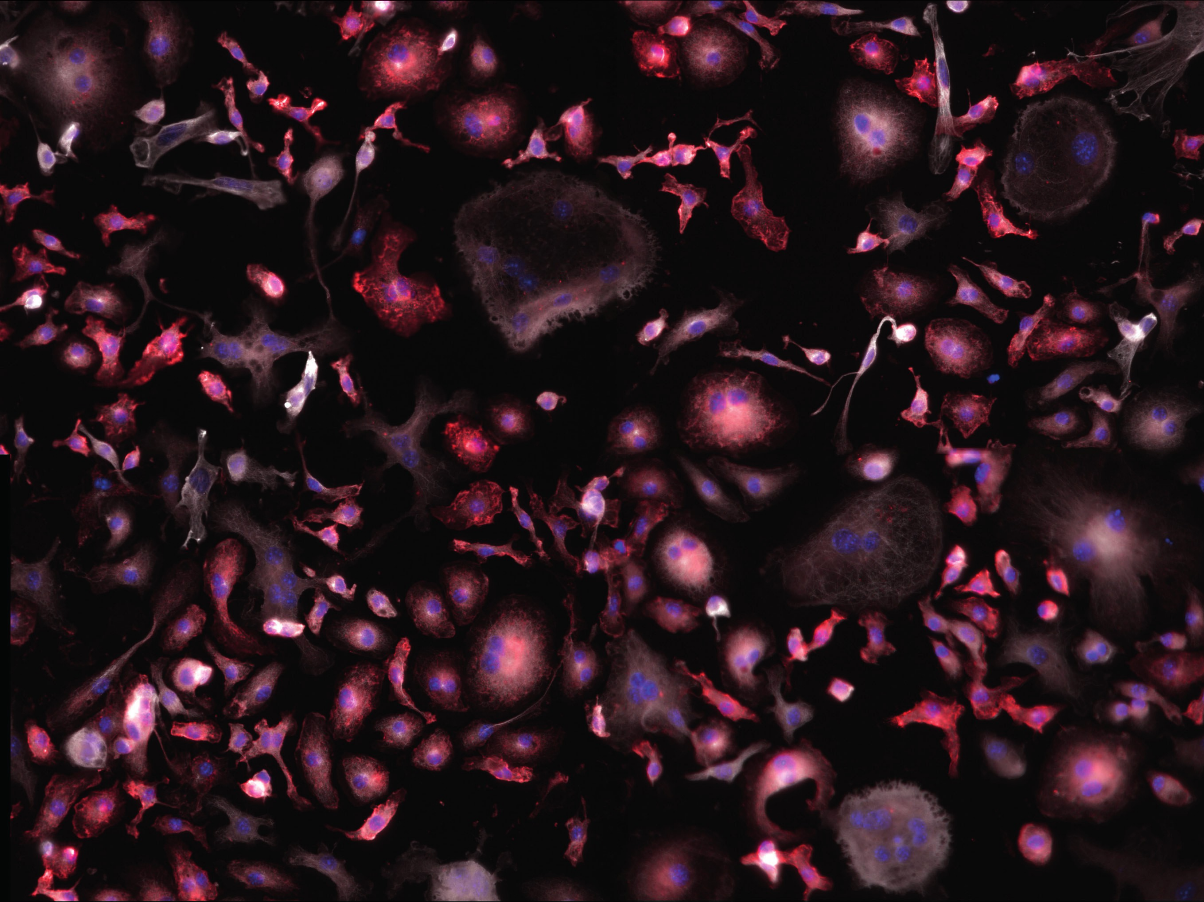
Osteoclasts are bone-resorbing cells in marrow whose pathology is implied in osteoporosis & rheumatoid arthritis. **We have built a system to segment & quantify osteoclasts in culture.**

© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

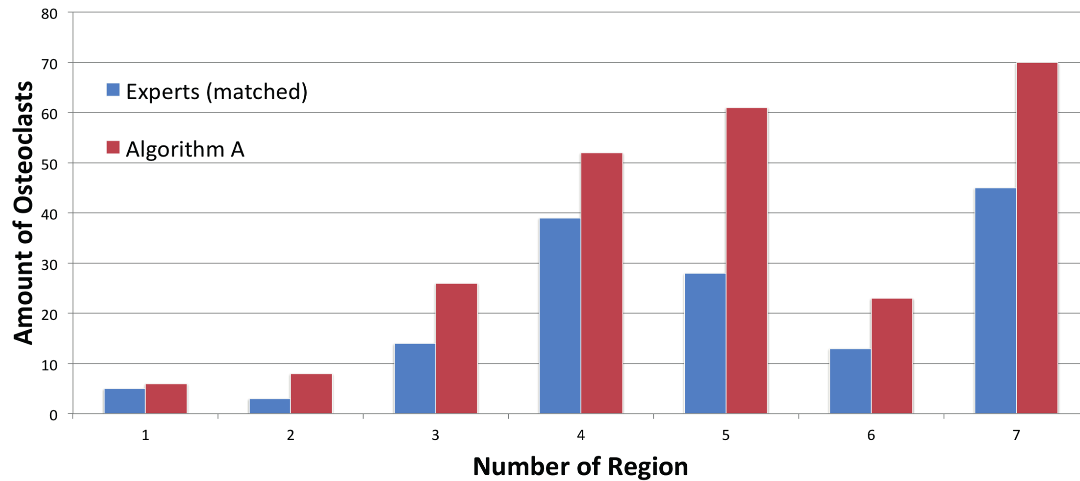
Osteoclast Characterization (2) - Sample



Triple Staining: white = cells, blue = nuclei, red = precursor / non-osteoclast

Detection works by counting nuclei (>=3) and computing red average area

Osteoclast Characterization (3) - Algorithm

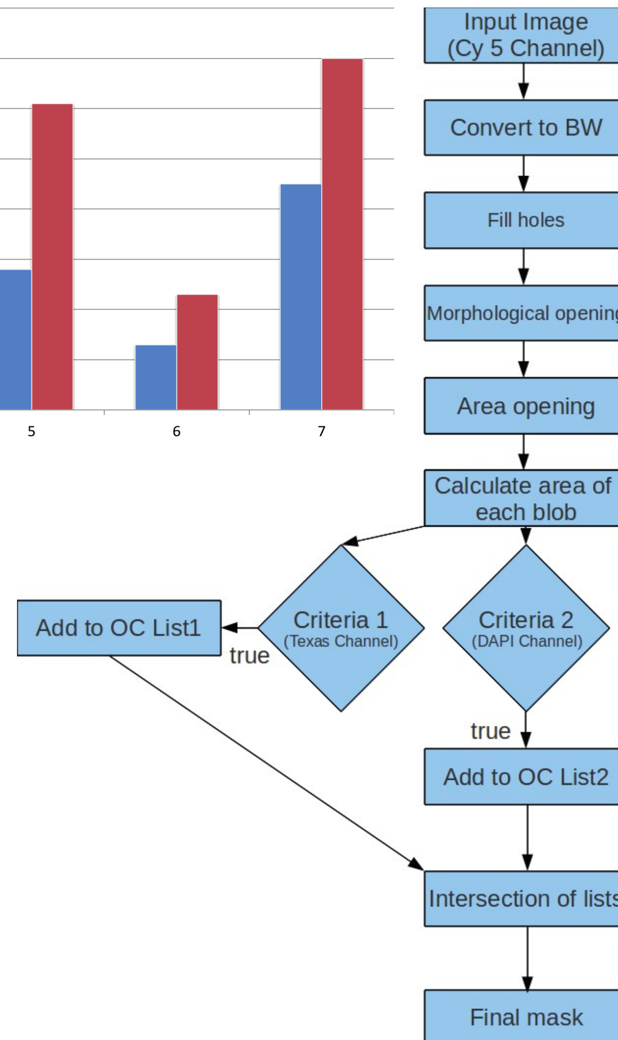


•Criterion 1 computes the ratio of the macrophage marker inside the cell versus the total area of the cell

$$\frac{amount(macrophage\ marker)}{area\ of\ cell} < T_1$$

•Criterion 2 counts the amount of DAPI staining inside the cell

$$amount(DAPI) > T_2$$



© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Osteoclast Characterization (4) - Current Work

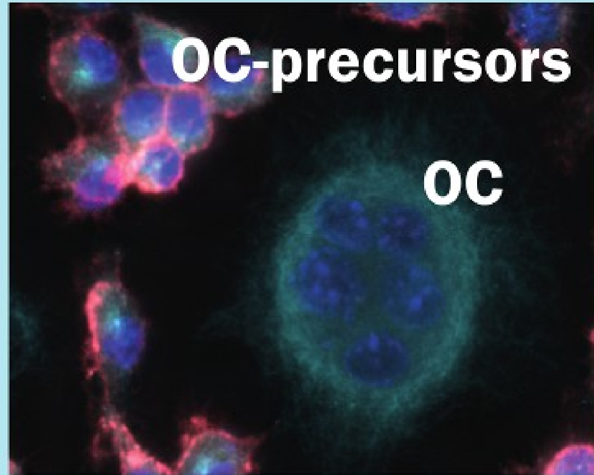


Figure 5

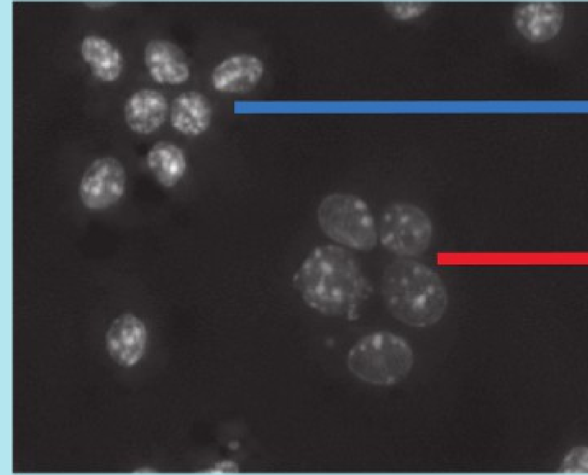
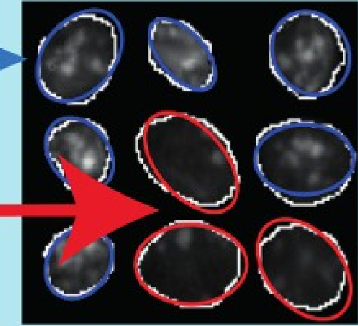
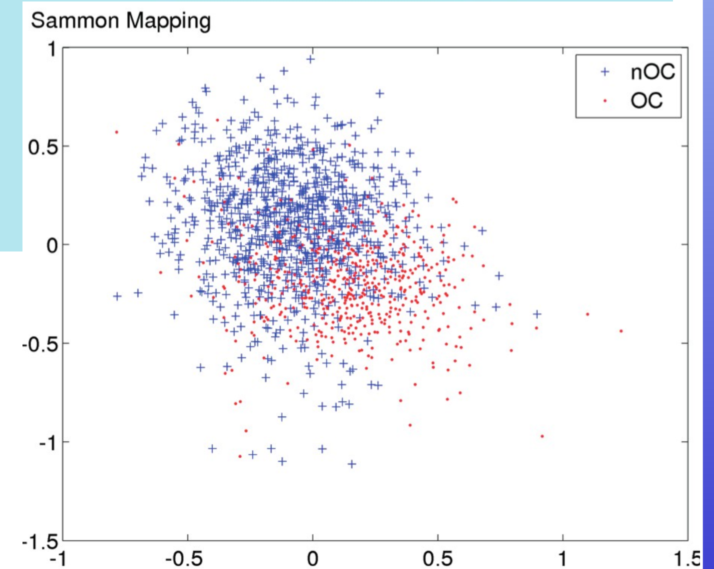


Figure 6





Shape and texture features computed from nuclei of osteoclasts (OC) and non-osteoclasts (nOC) allow to determine class with some confidence (Logistic regression 10CV: 83% accuracy, 75% on holdout test sample) May be combined with current system for additional precision.

© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

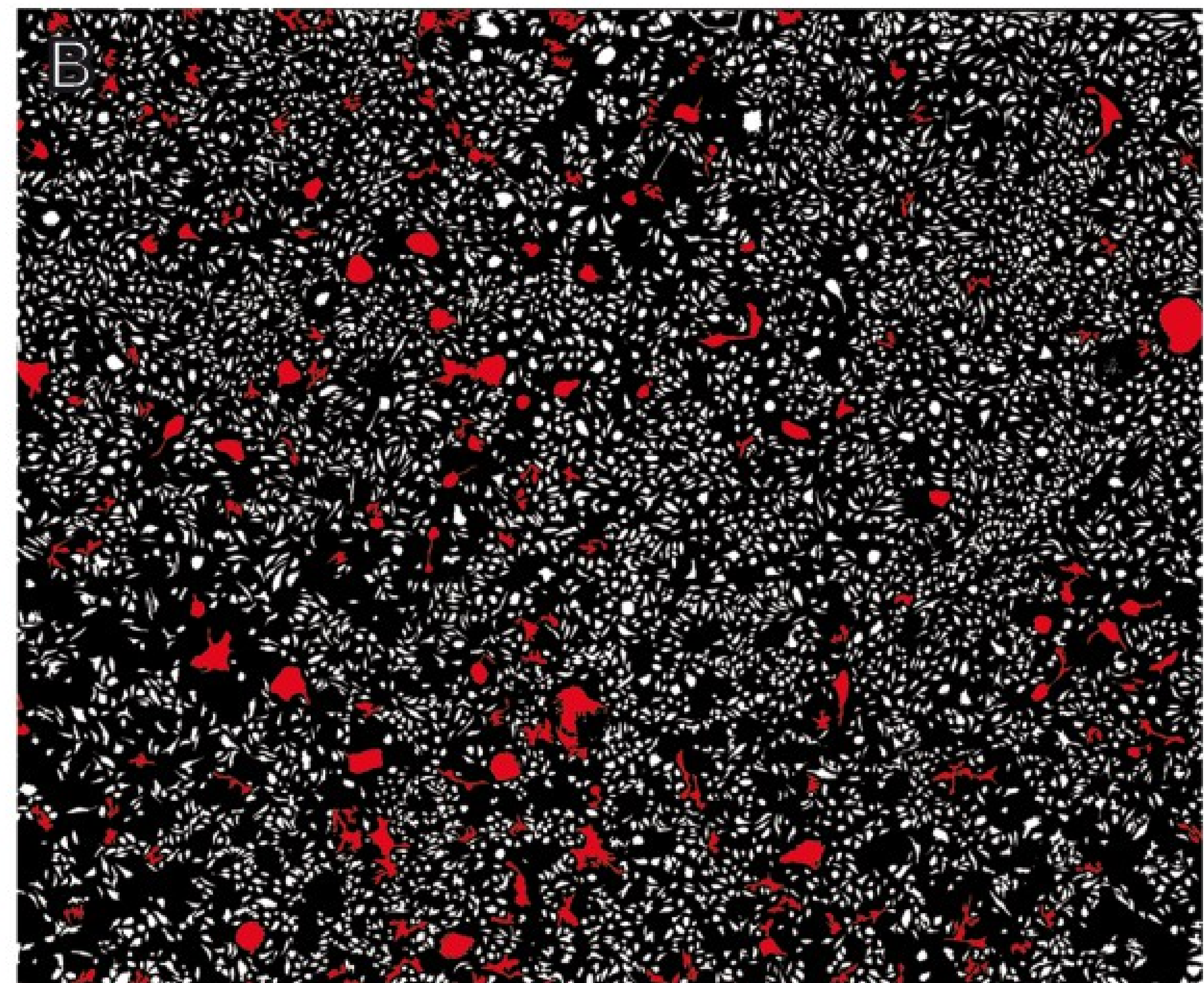
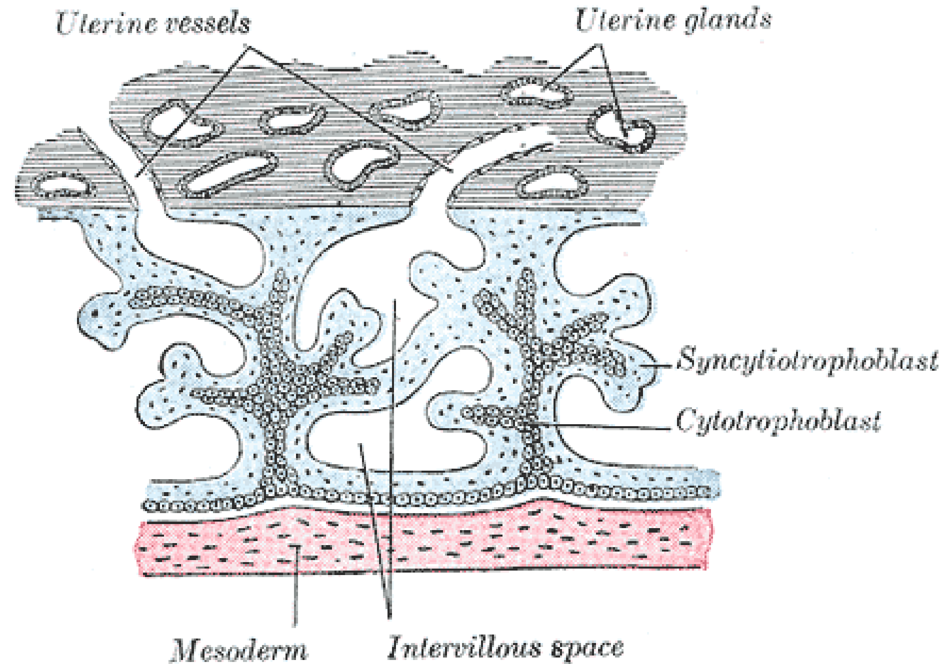


Figure 4: Image analysis of a whole sample-region consisting of 100 (10 x 10) FOVs. A: Immunofluorescence image. B: Result of cell segemntation and analysis: non-OC are marked in white, OC are marked in red.

Syncytiotrophoblast (1)

Primary chorionic villi



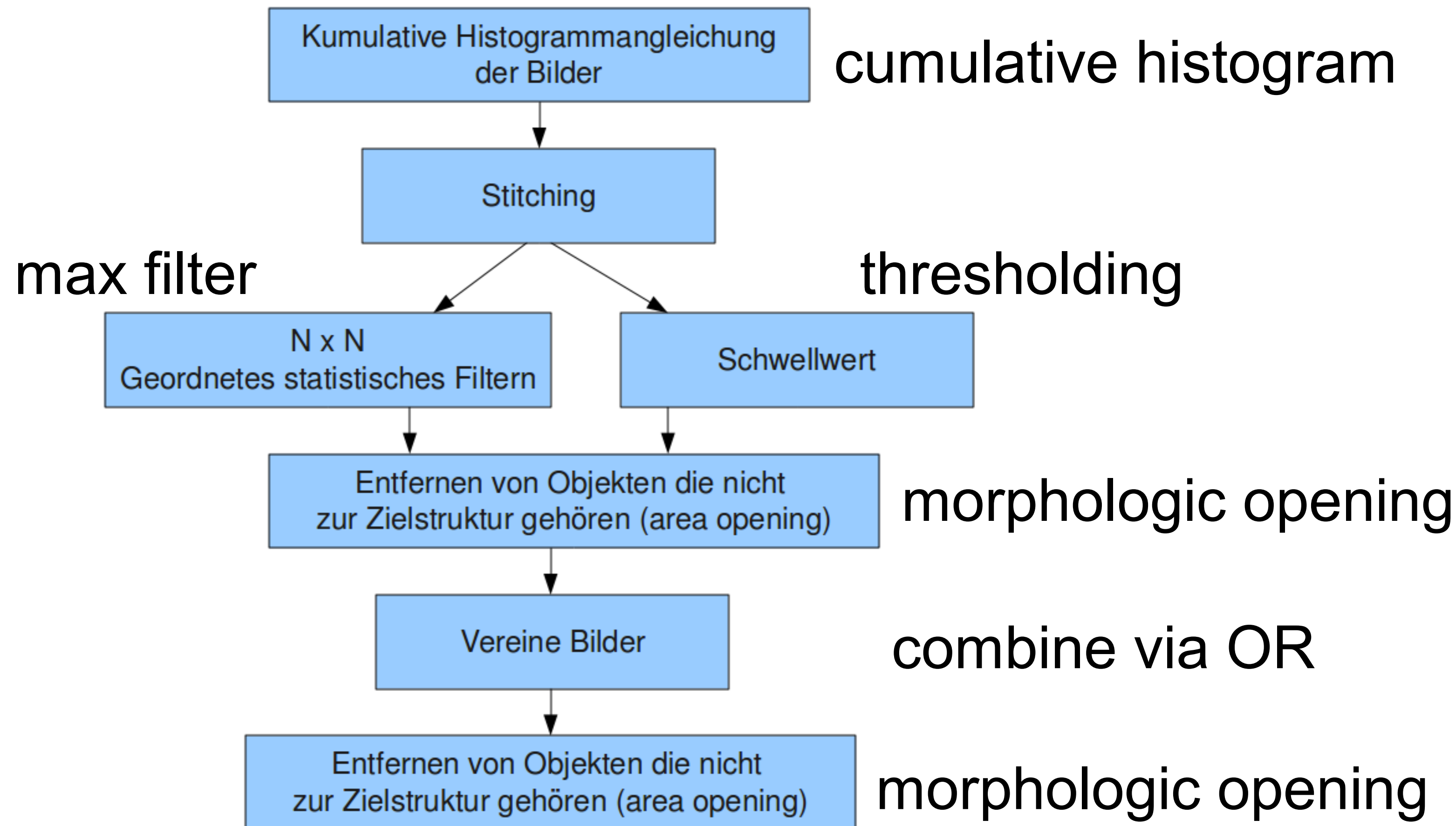
Syncytiotrophoblasts are multinucleated cells within the placenta of embryos at the surface of chorionic villi. Chorionic villi are part of the border between maternal and fetal blood during pregnancy. **We have built a system to segment villi & syncytiotrophoblast and applied it to protein quantification of Receptor for Advanced Glycated End products(RAGE)**

© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Syncytiotrophoblast (2) - Algorithm



Exhaustive Parameter-Optimization

&

Around 8,100 sets of 7 parameters were tested.

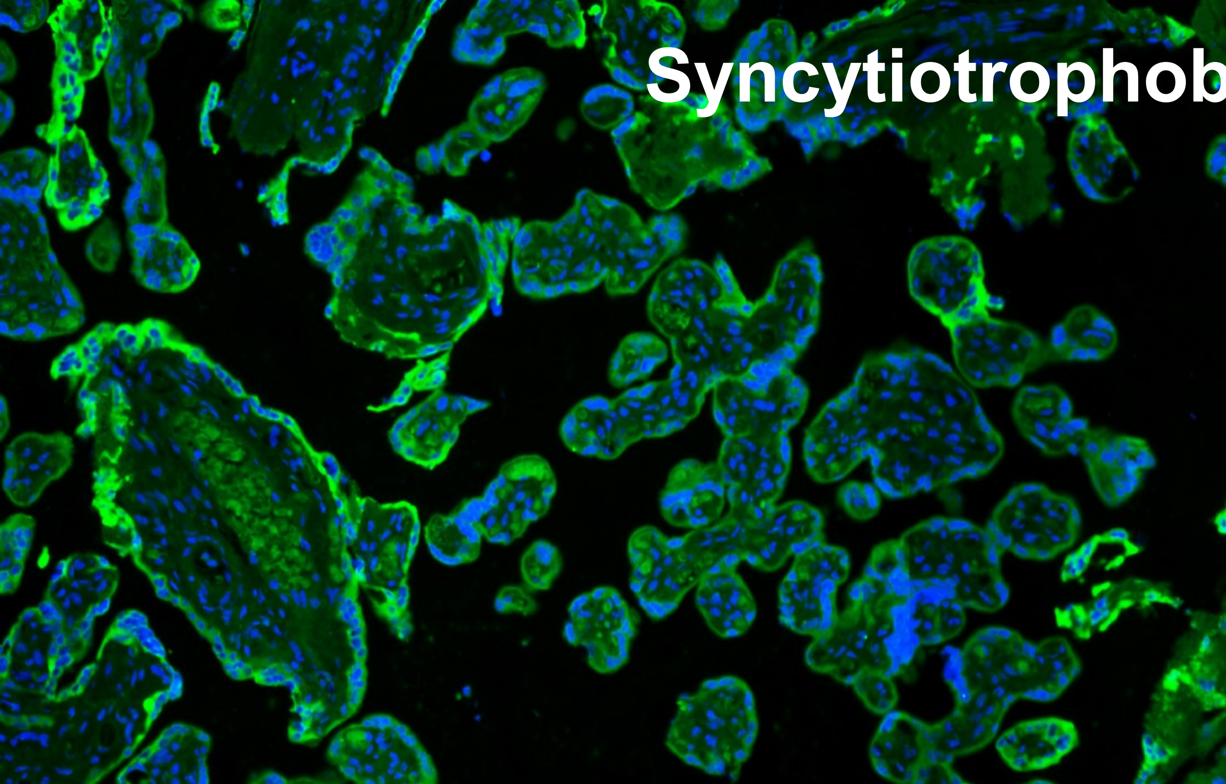
Improved F1 from 0.78 to 0.93

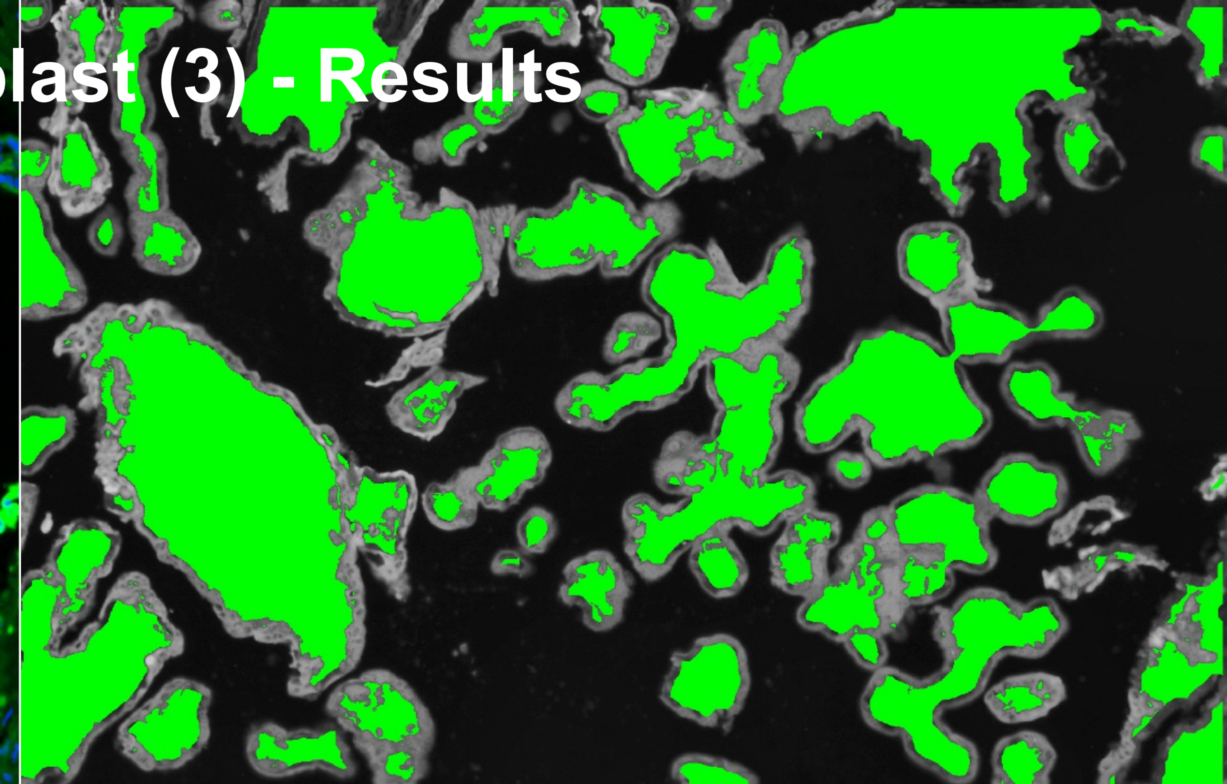
© 2012

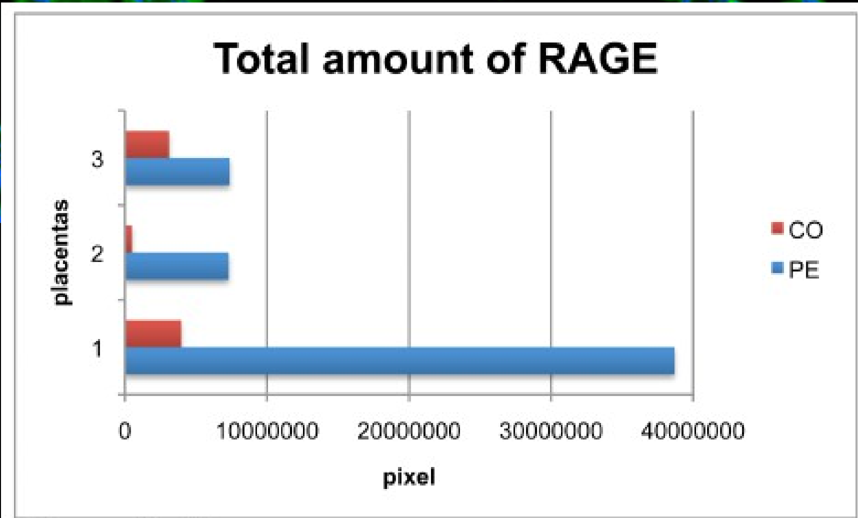
SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Syncytiotrophoblast (3) - Results







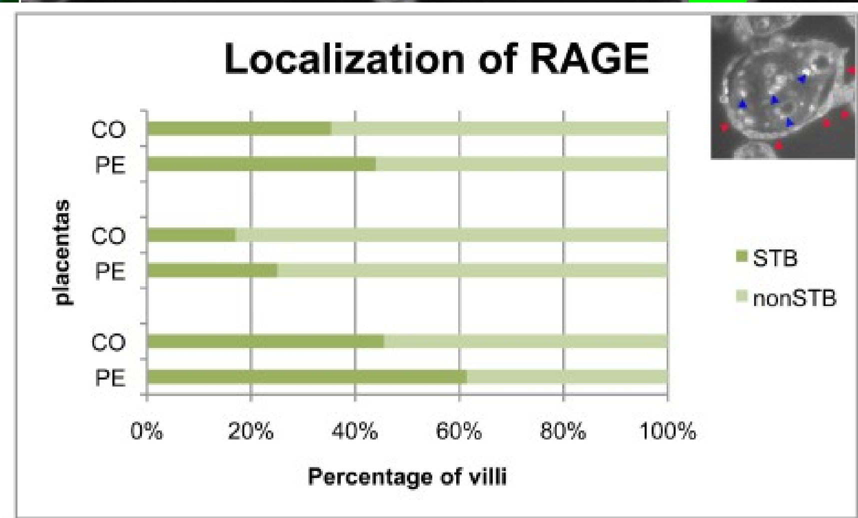
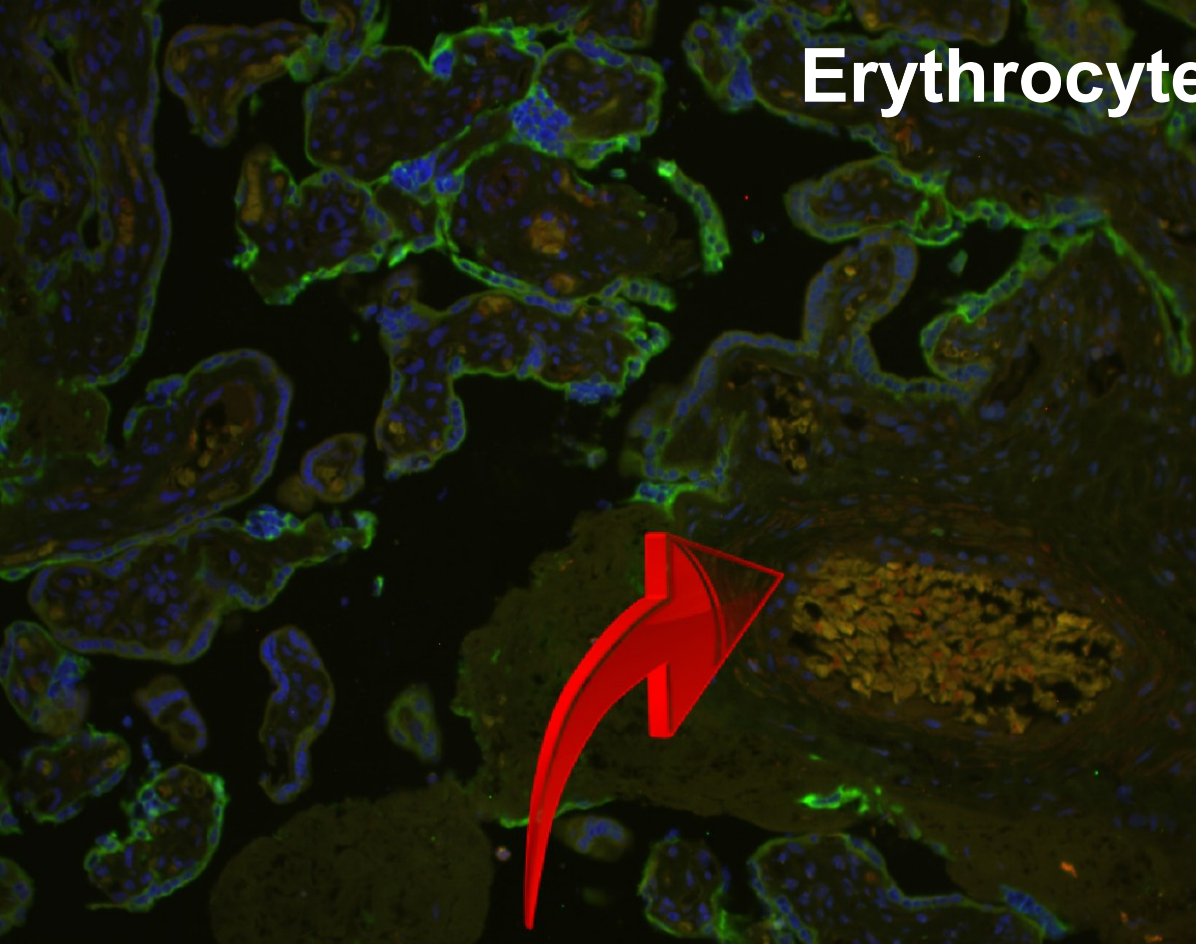
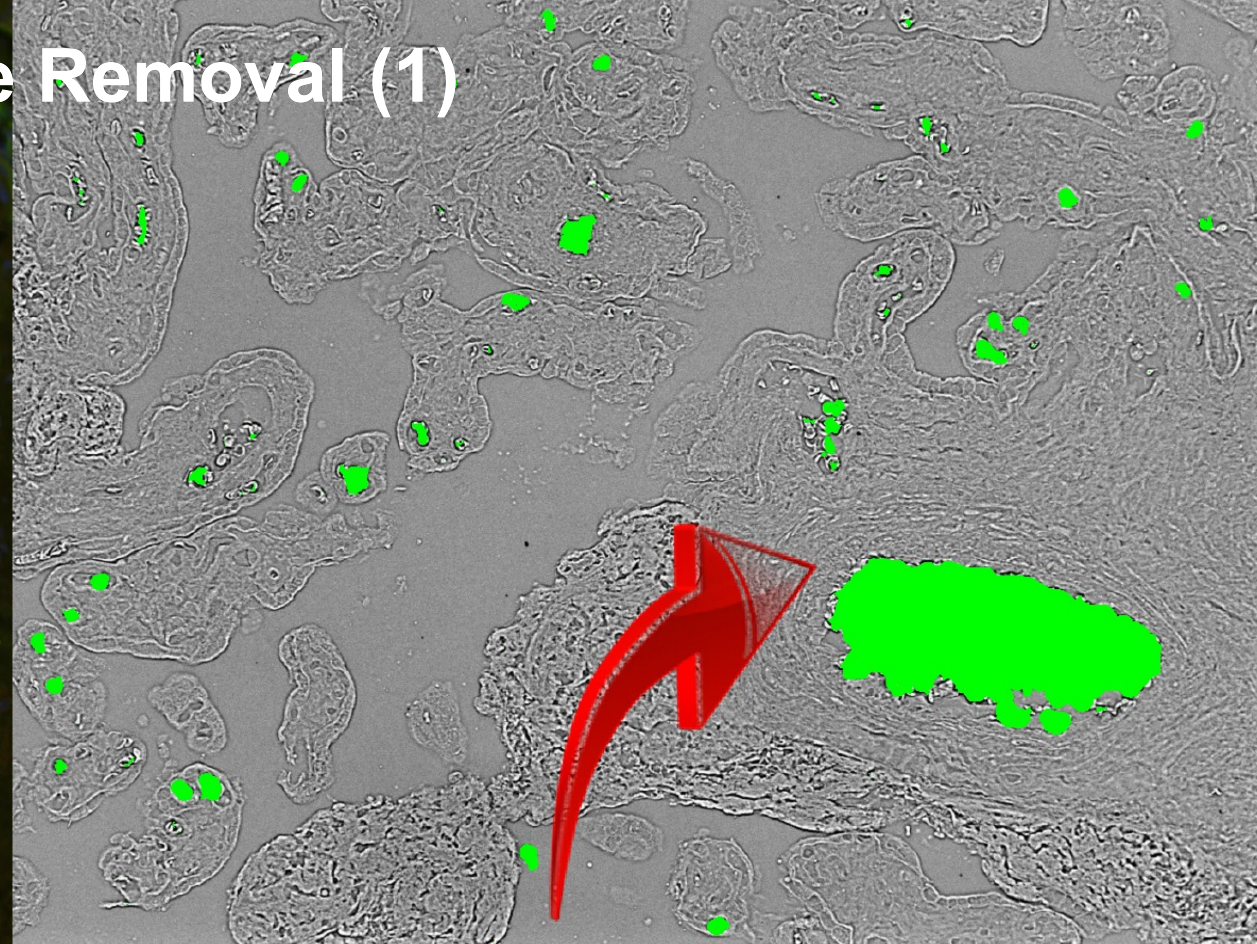


Figure 8-10

The total amount of RAGE proteins quantified in PE and CO placentas is shown in Figure 8, respectively localization within the villi in Figure 10. RAGE can be found in STB (red arrowheads) or other cell types (blue arrowheads) visualized in Figure 9.

Erythrocyte Removal (1)





Erythrocytes have high autofluorescence and may lead to noisy immunofluorescent measurements. **Using only ground-truth data, we „taught" the computer to remove erythrocytes from images ([Viola & Jones, 2001] = OpenCV Haartraining)**

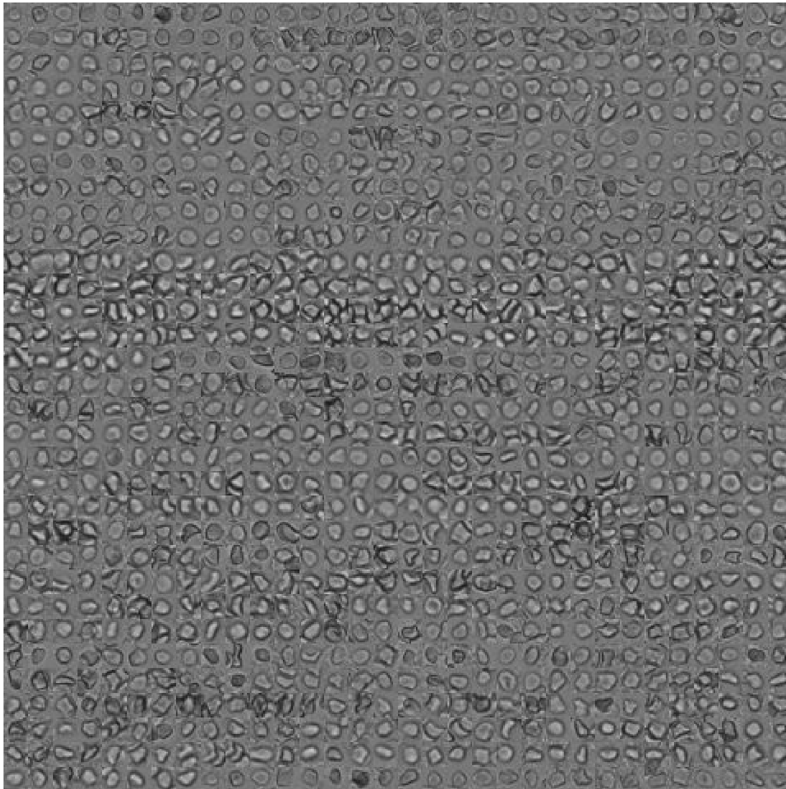
Essential for RAGE analysis from previous slide!

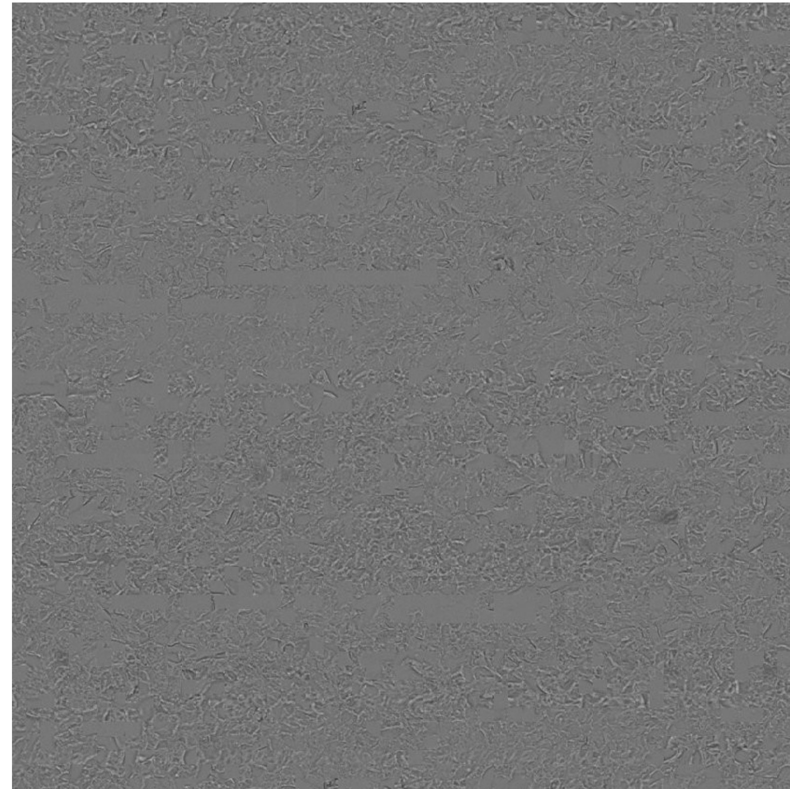
Erythrocyte Removal (2) - Training

Training Data

>4000 erythrocytes have been marked up manually as well as...

... *>2000* regions containing no erythrocytes (tissue, background)





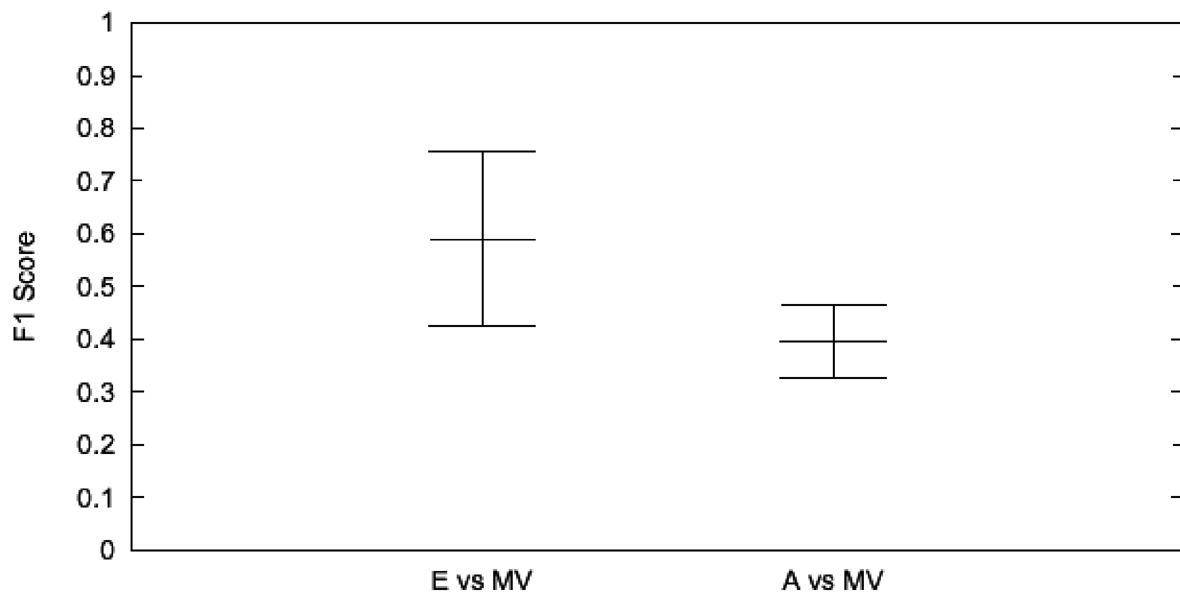
&

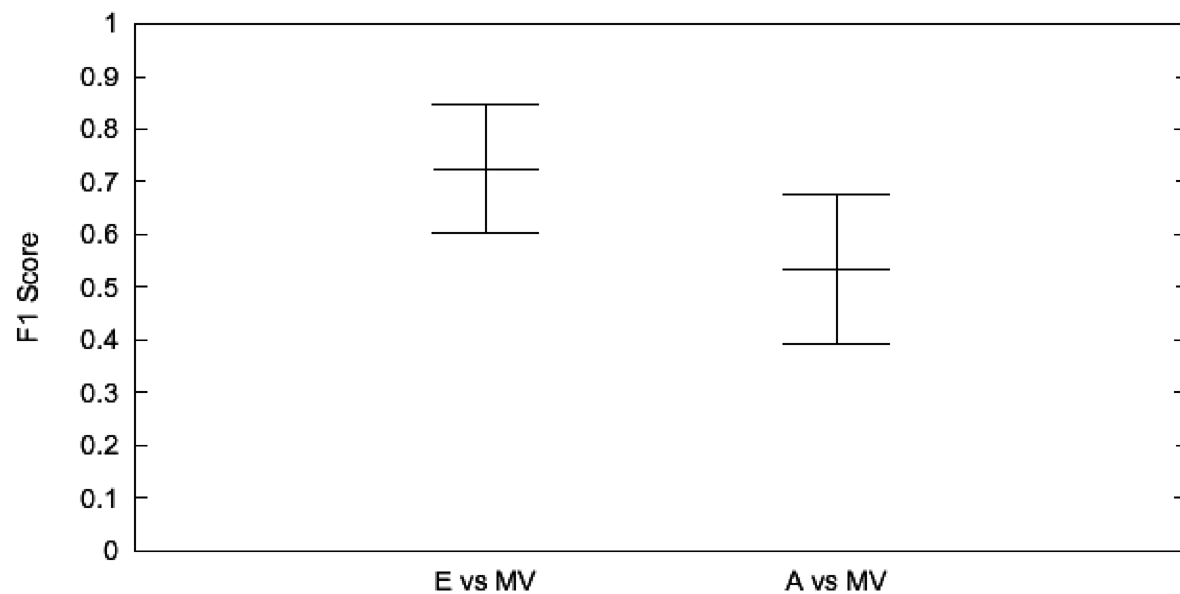
- Exhaustive Haar-Train-Param.Opt (1.8 million masks)
- >=2 Experts for each FOV, many train data variants
- Extend Haar-Test with conf. values
- Run four times with lossless k*90° rotations...

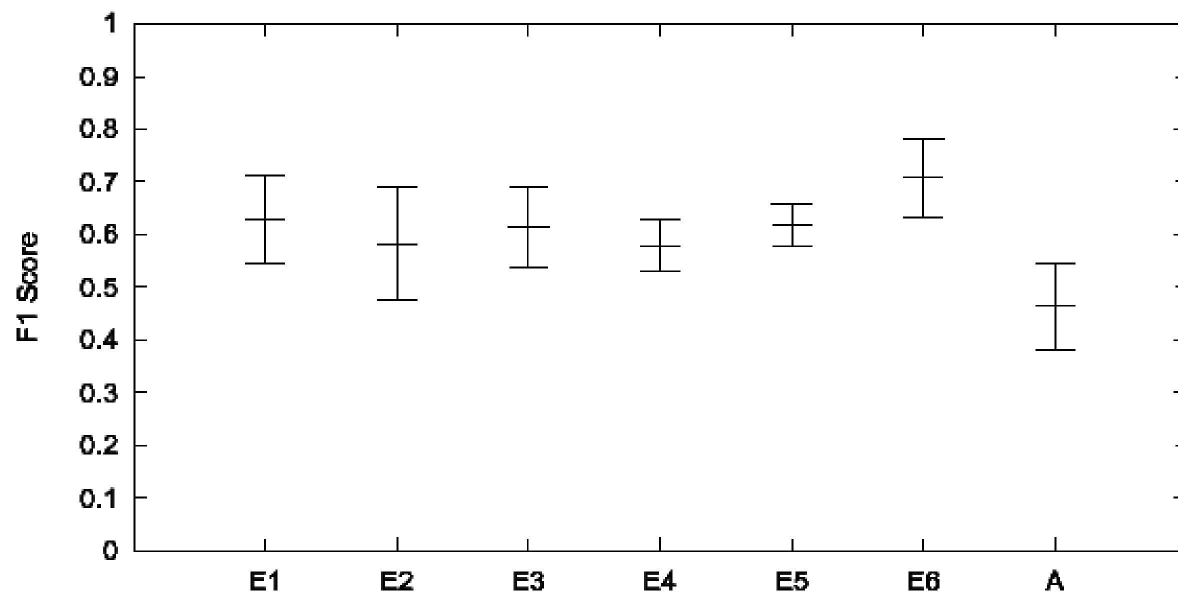
© 2012

SEEWALD SOLUTIONS

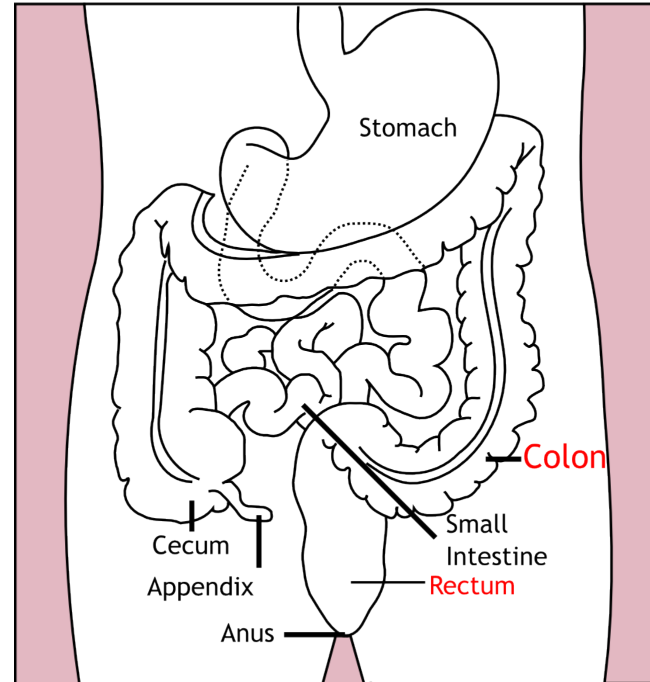
alex@seewald.at
www.seewald.at


Colon
F1 Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
E vs MV
A vs MV


Placenta
F1 Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
E vs MV
A vs MV


Performance (F1 Score) Algo vs 6 experts on one FOV
F1 Score
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
E1
E2
E3
E4
E5
E6
A

Colorectal Cancer (1)

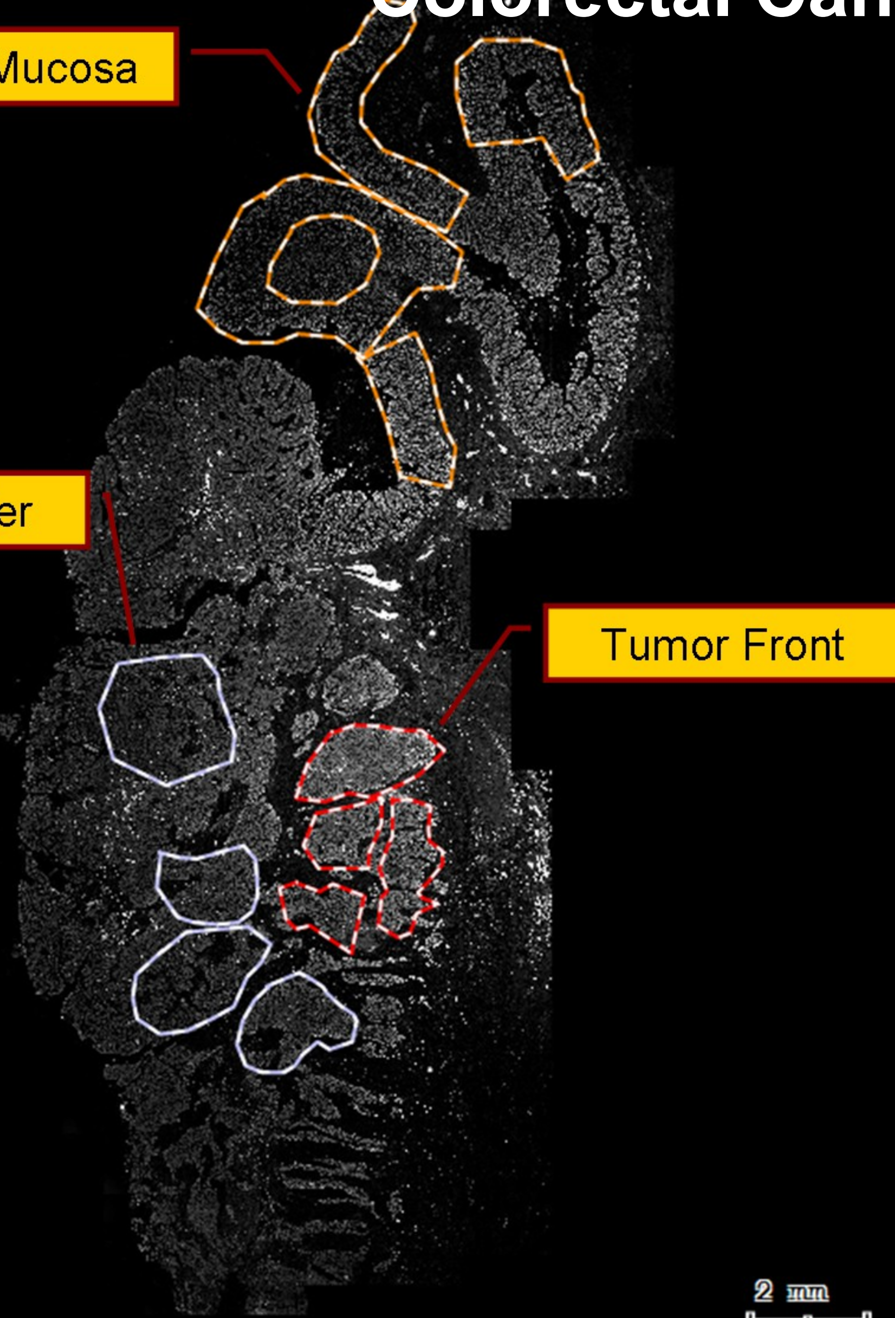


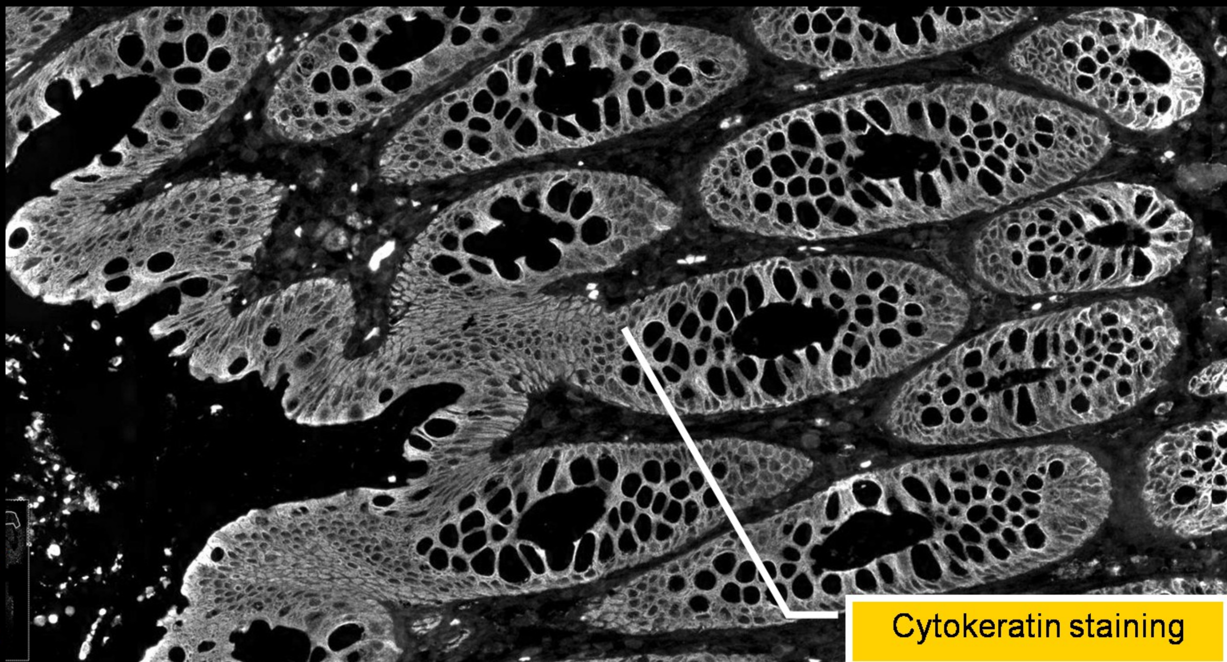
Colorectal cancer (CRC) is the second most frequent cause of cancer-related mortality in the Western world. Increasing evidence supports the role of the tumor microenvironment in influencing tumor progression (e.g. epithelium-to-stroma ratio) **We built a system to identify stroma, epithelium and lumen in colon tissue as a first step towards automated diagnosis and prognosis.**

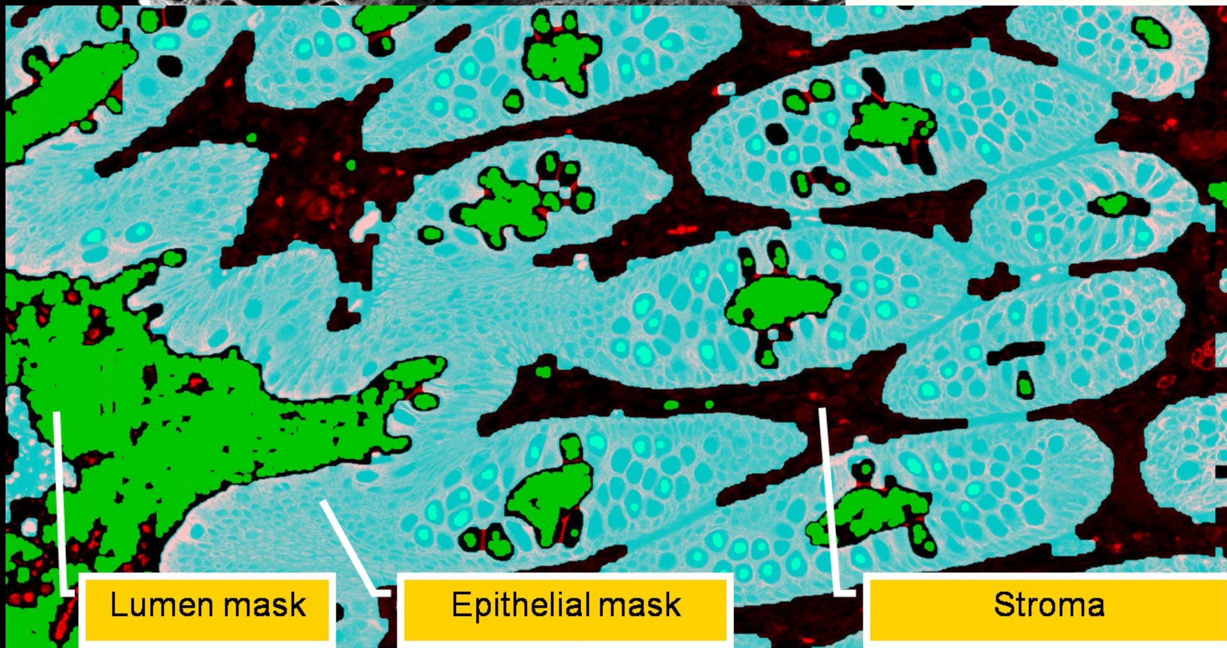
© 2012

alex@seewald.at
www.seewald.at

Colorectal Cancer (2) - Sample







Colorectal Cancer (3) - Algorithm

(Non-)Lumen / Tissue Mask

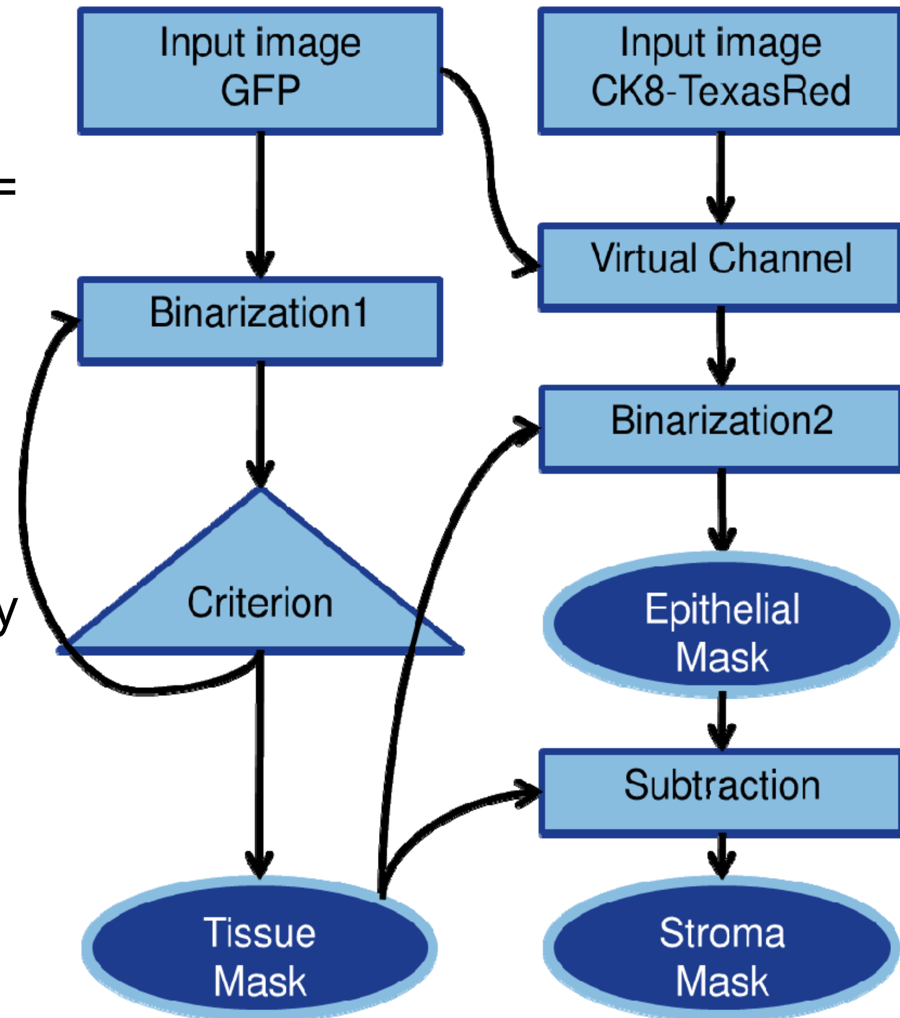
- Input: GFP autofluorescence channel only
- Adaptive binarization(1) with success criterion = low variance of Lumen area vs. Non-Lumen area

Epithelial Mask (CK8 staining)

- Combine all channels into virtual channel
- Non-adaptive Binarization(2) within the previously computed Tissue Mask

Stroma Mask

- Subtract Epithelial Mask from Tissue Mask



© 2012
SEEWALD SOLUTIONS
alex@seewald.at
www.seewald.at

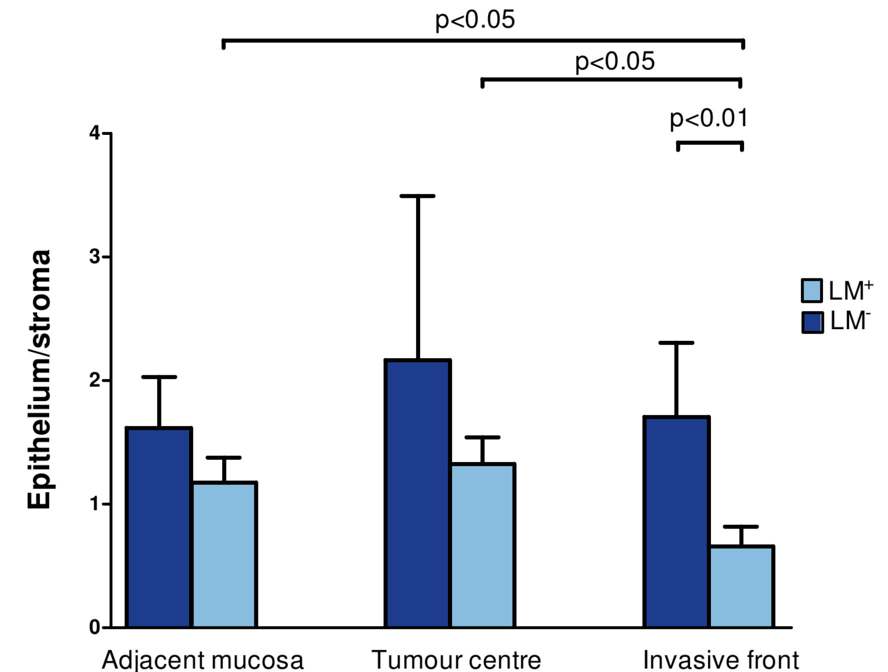
Colorectal Cancer (4) - Results

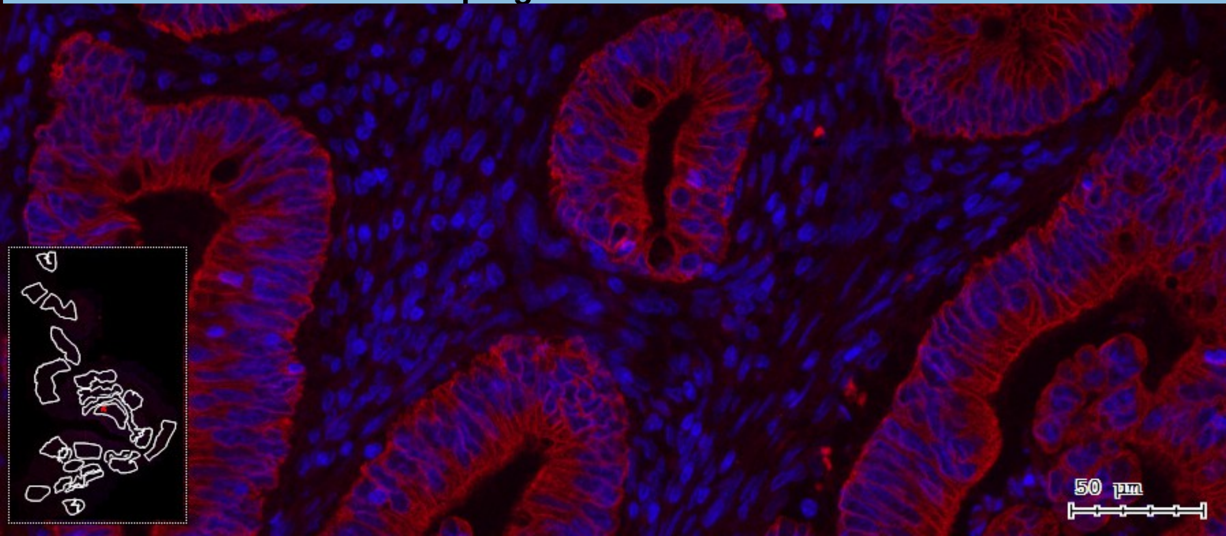
Colorectal Cancer (4) - Results

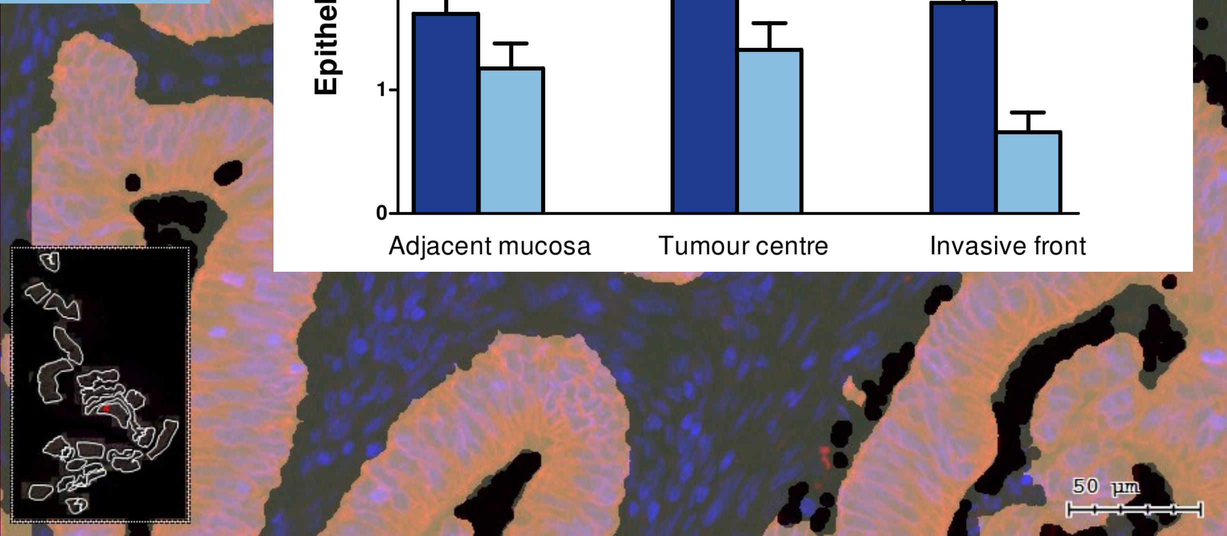
CONCLUSIONS AND OUTLOOK

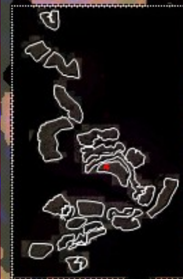
1) The strong reduction of the epithelium/stroma ratio at the invasive front of colon cancer tissue of patients with liver metastasis confirms the benefit to use this parameter as an additional information in CRC diagnosis.
2) Our results are in agreement with previous studies that used visual methods.
3) The implemented software can be used for detailed morphometric analysis of CRC tissue, as well as for the quantitative evaluation of markers expressed in the epithelial or stromal area for a better characterization of epithelium-stroma interaction in tumour progression.

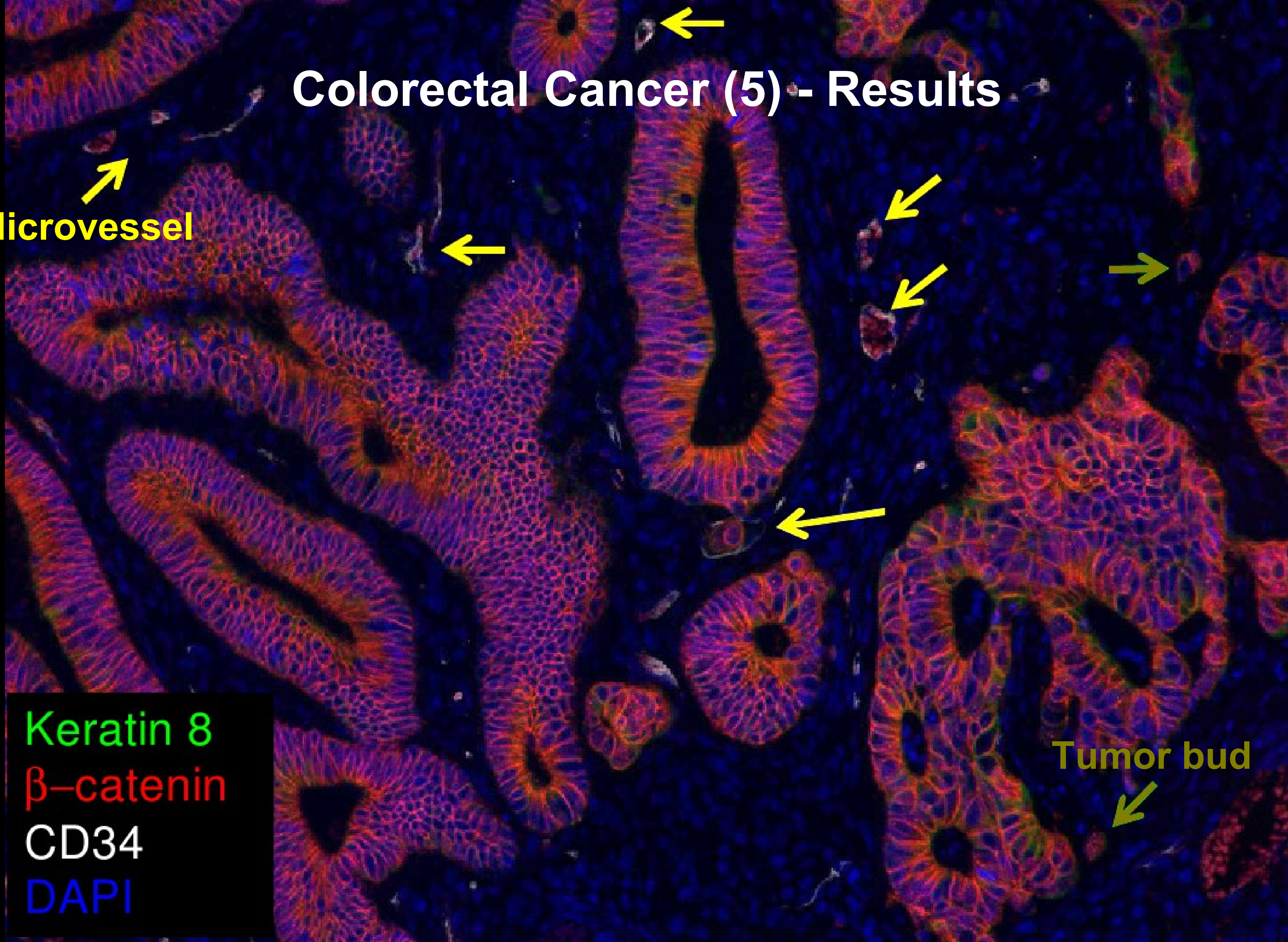
Figure 3 – Epithelium/stroma ratio in different sub-regions in colorectal cancer patients with and without liver metastasis











Colorectal Cancer (5) - Results
Microvessel
Tumor bud
Keratin 8
β–catenin
CD34
DAPI

Colorectal Cancer (5) - Results

Microvessel

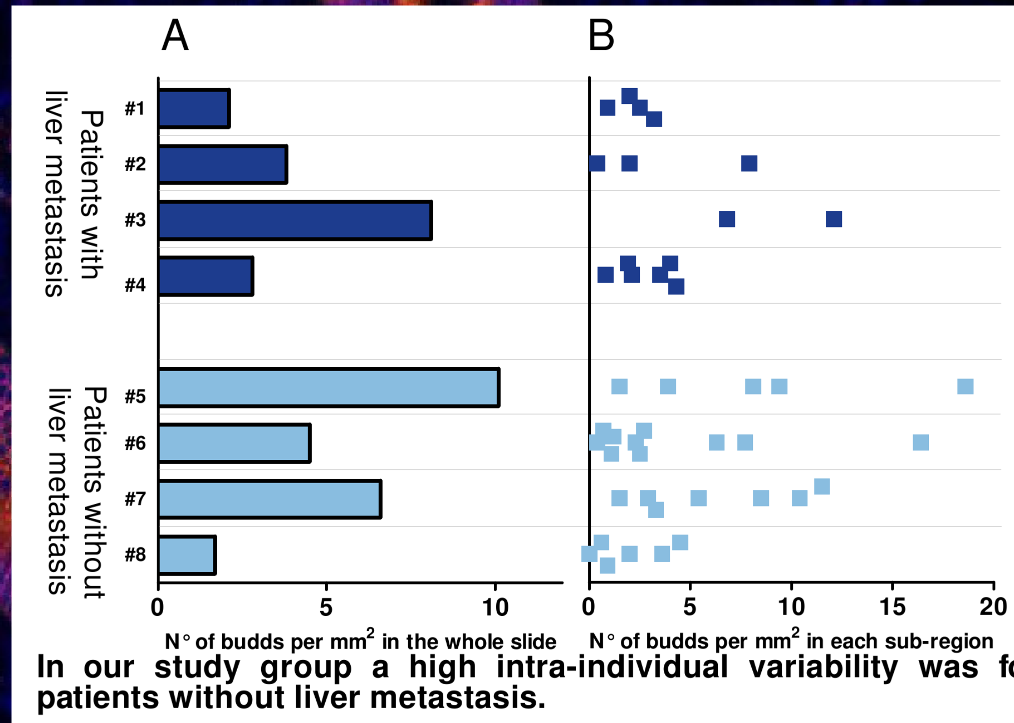


Figure 4. Intra-tumoral budding was manually assessed in tissue specimens of 4 colorectal cancer patients with liver metastasis (#1 - #4) and 4 without liver metastasis (#5 - #8). Results are presented as buds per mm^2. **(A)** The total number of buds counted in each specimen was related to the overall tumor center area. **(B)** Each tumor center sub-region is represented by a square.

In our study group a high intra-individual variability was found, more pronounced in patients without liver metastasis.

CONCLUSION AND FUTURE DIRECTIONS

Our data suggest no correlation between intra-tumoral budding and the presence of liver metastasis in patients with colorectal cancer grade 2. This might me due to the limited number of patients. Therefore, we propose assessing intra-tumoral budding in a larger cohort of patients. Furthermore, we aim to evaluate budding in relation to other factors such as distance to blood and/or lymphatic vessels.

Keratin 8
β–catenin
CD34
DAPI

C. Elegans Protein Localization (1)

Seewald AK, Cypser J, Mendenhall A, Johnson T (2010) Quantifying Phenotypic Variation in Isogenic Caenorhabditis elegans Expressing Phsp-16.2::gfp by Clustering 2D Expression Patterns, PLoS ONE 5(7): e11426. doi:10.1371/journal.pone.0011426.

Quantifying Phenotypic Variation...

Analyzing changes in appearance / phenotype...

in Isogenic Caenorhabditis elegans...

in small nematodes (worms) which all have the same genetic code (clones)

Expressing Phsp-16.2::gfp...

which express a GFP reporter that binds to heatshock protein 16 (transgenic)

by Clustering 2D Expression Patterns

by extracting 2D expression patterns that are independent of worm pose AND clustering these patterns using hierarchical clustering methods.

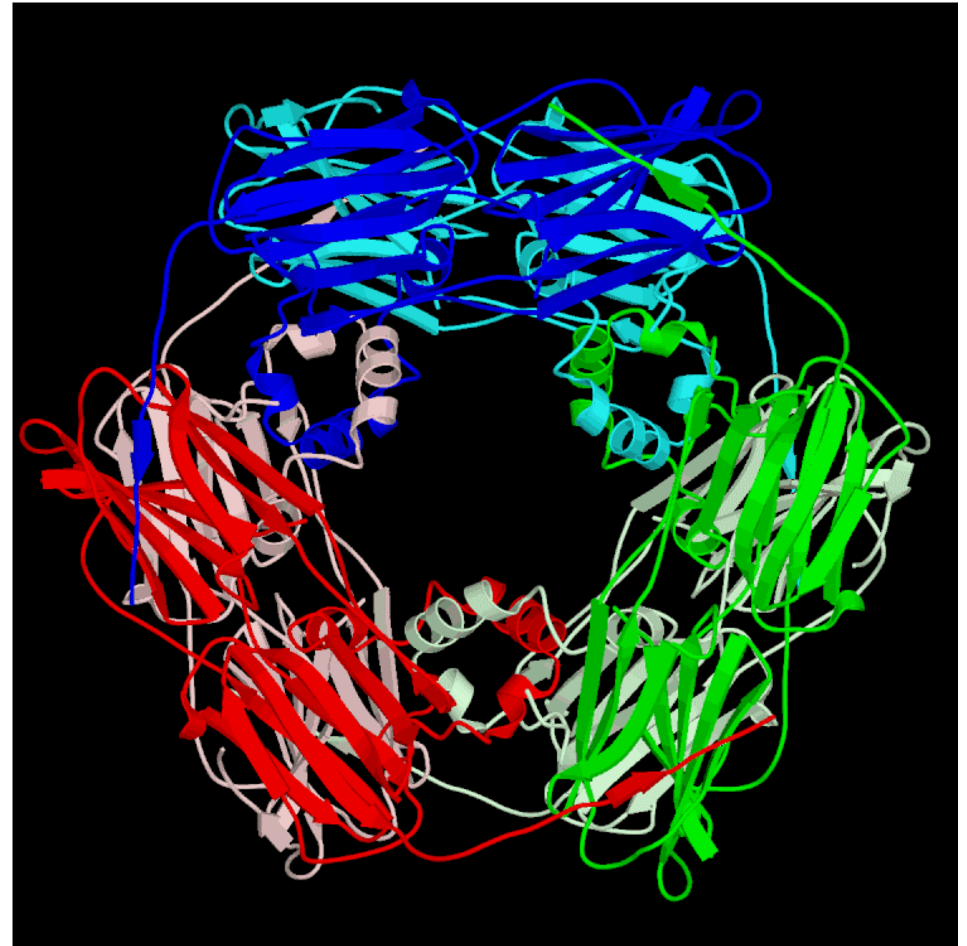
© 2012

alex@seewald.at
www.seewald.at

C. Elegans Protein Localization (2)

Heat Shock Protein 16 – Increases expression (also) when organism is exposed to high temperatures

- HSP are named by molecular weight (=16kD). Expressed in intestine and pharynx. Induced in response to heat shock or other environmental stresses.
- Interacts with intra-cellular human beta amyloid peptide (Alzheimer plagues)
- High expression correlates with worm longevity acc. to earlier studies.



© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

C. Elegans Protein Localization (3)

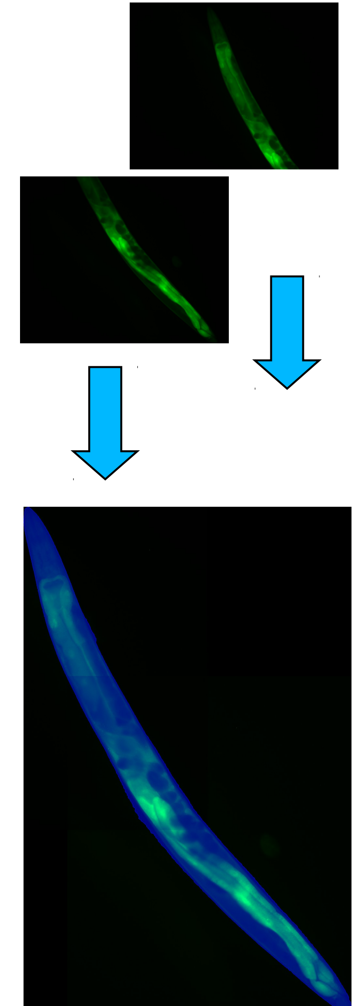
Merging anterior/posterior image & extracting worm

(pixelClassification)

- Machine learning from manually tagged sample images (i.e. „Ground Truth“)
- Threshold optimization by testing minimum circularity and area of largest blob
- Closure (erode, dilate), fill internal holes with circularity below threshold
- Heuristic search for breaks in contour, which are repaired with straight lines and filled on the inside

(meshAB)

- Image correlation coefficient for combining head and tail images (~ simplified stitching)



© 2012
alex@seewald.at
www.seewald.at

C. Elegans Protein Localization (4)

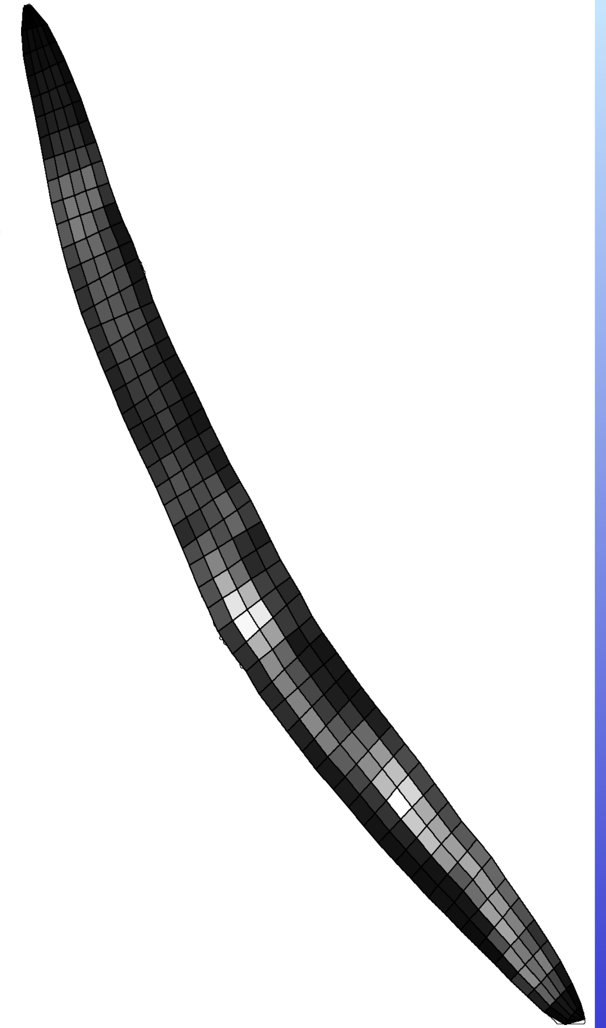
Computing 2D expression patterns

(meshAB)

- Computing head/tail/vulva position from worm pose via heuristics (~ 50%)
- Manual markup in remaining cases

(sampleCE)

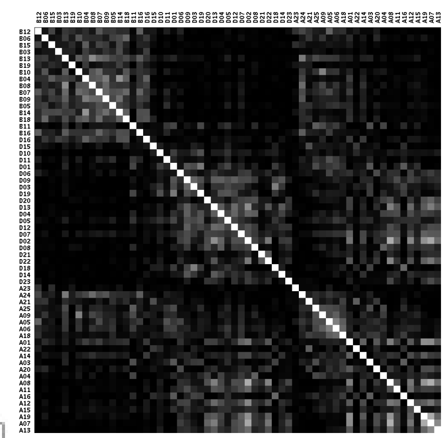
- Worm backbone via „thinning“
- Search left/right from backbone in perpendicular direction to local curvature for worm border
- Split left/right and top/bottom into the desired number of tiles
- Compute average GFP intensity per tile



© 2012

alex@seewald.at
www.seewald.at

C. Elegans Protein Localization (5)



Known: Bright worms live longer than dim ones

- Even when discounting brightness, bright worms show distinct expression patterns (currently under investigation)
- Dim worms show two clusters of activation patterns

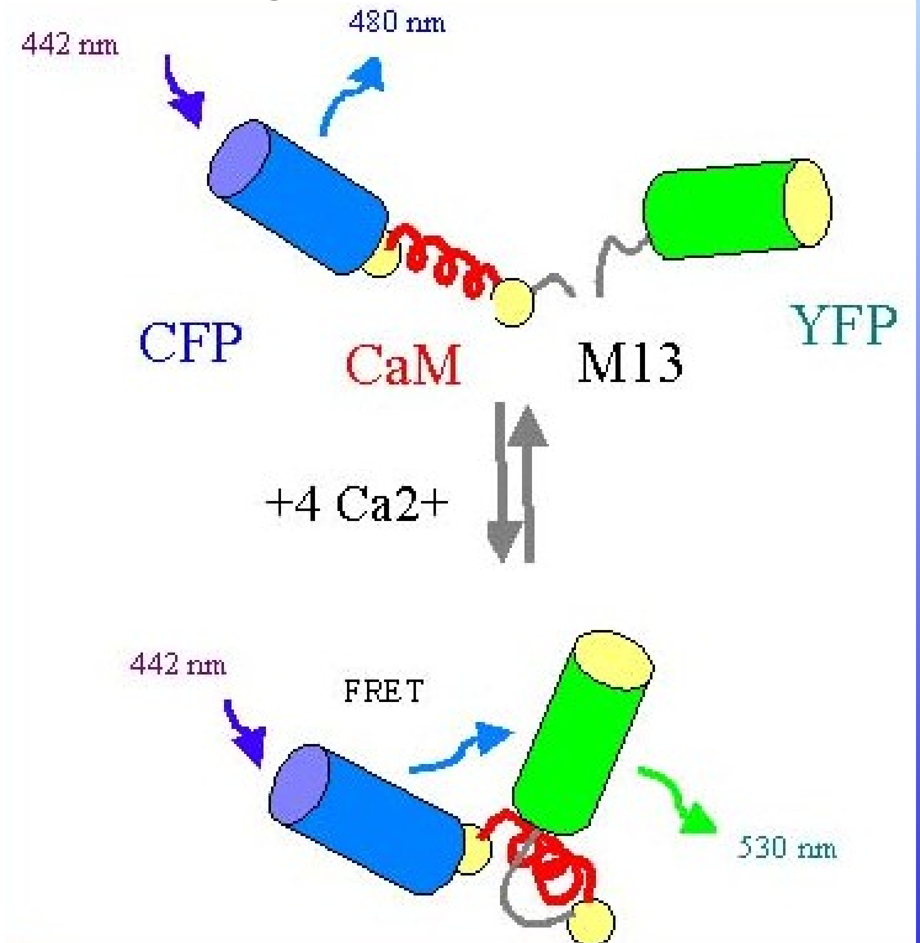
Watching C. Elegans Think

Cameleon: measures Ca2+ level = nerve cell activity

CFP emits a 480nm photon on excitation with 442nm.

High Ca2+ concentrations lead to conformation changes and the photon is absorbed by YFP and re-emitted as 530nm.

Proportion between 480nm and 530nm response used as signal for Ca2+ level.



© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

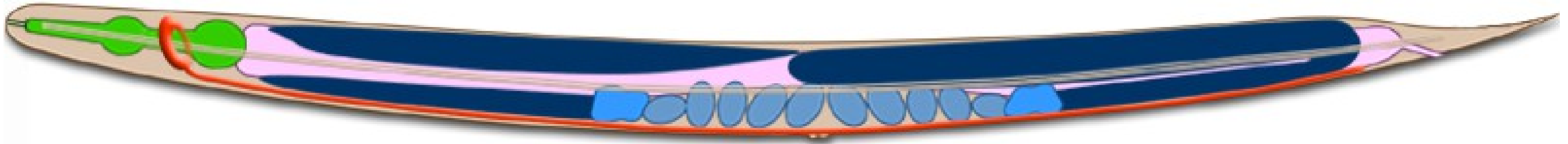
Video 2 - Worm Dedistorting in Real Time

© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

C. Elegans Protein Localization (6) - Future Work



- Bottleneck is image acquisition – each worm has to be taken from culture medium, anesthesized, cleaned and imaged (ca. 30min per worm)
- Resolution is too coarse for observing single cells

→ *Culturing worms on chamber-slides, using slide-based microscopy & automated imaging*

- Lots of problems with different microscope settings, air bubbles, finetuning,... one common major problem in designing robust image analysis systems

→ *„Closed-loop“ system (microscope, moveable slide and image analysis - coupled system)*

© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Conclusion

Designing Image Analysis Systems is tough and takes time...

- Four systems designed within three year FFG Bridge project
  2 PhD students, 1 CompSci researcher, 3 biological res., ~7 additional students

- Obtaining useful images and reliable ground-truth data is the key!
  Sometimes tasks are ill-defined and/or experts disagree on classification

- Machine Learning techniques can speed up implementation significantly but are currently applicable only to restricted imaging tasks (e.g. *Erythrocyte Removal*)

- Parameter optimization & algorithm finetuning with ground truth data can dramatically improve even simple systems.

Thanks for listening!

© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at

Questions (1)

We saw that automated image processing system are able to replicate expert human performance reasonably well in some areas. However, experts do not always agree. On the other hand it is probably not advisable to let an automated system take life-or-death decisions, giving that so many factors outside the control of the system are relevant (e.g. image quality, antibody purity, staining expertise...)

Where would you draw the line between relying on automated, calibrated systems and relying on holistic, intuitive human expertise?

a) checking your biochemical pathways and suggesting lifestyle changes (e.g. less sugar, more carbohydrates, more sports)

b) an operation to remove a possibly (pre-)cancerous mole

c) an operation to remove a kidney or parts of the colon

d) determining who gets which heart for a transplant

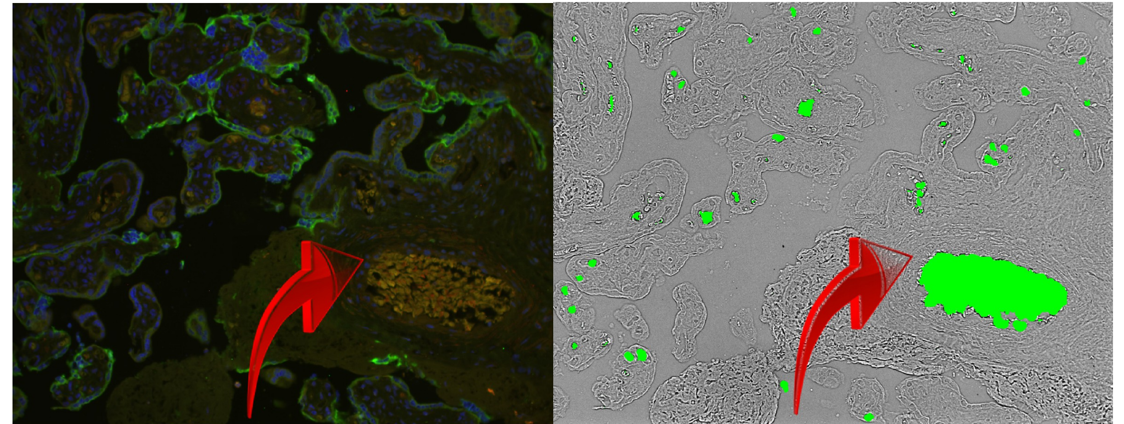
© 2012

alex@seewald.at
www.seewald.at

Questions (2)

Of the automated image analysis systems which were explained, most were of the classical kind and were designed stepwise by a human expert. However, one system - *Erythrocyte Removal* - learned only from ground truth data and did not involving programming in any sense.

Assume two systems which solve the same image analysis task - one a classical program, the other one learning from ground truth data - and that both solve this task equally well. Which one would you prefer and why?



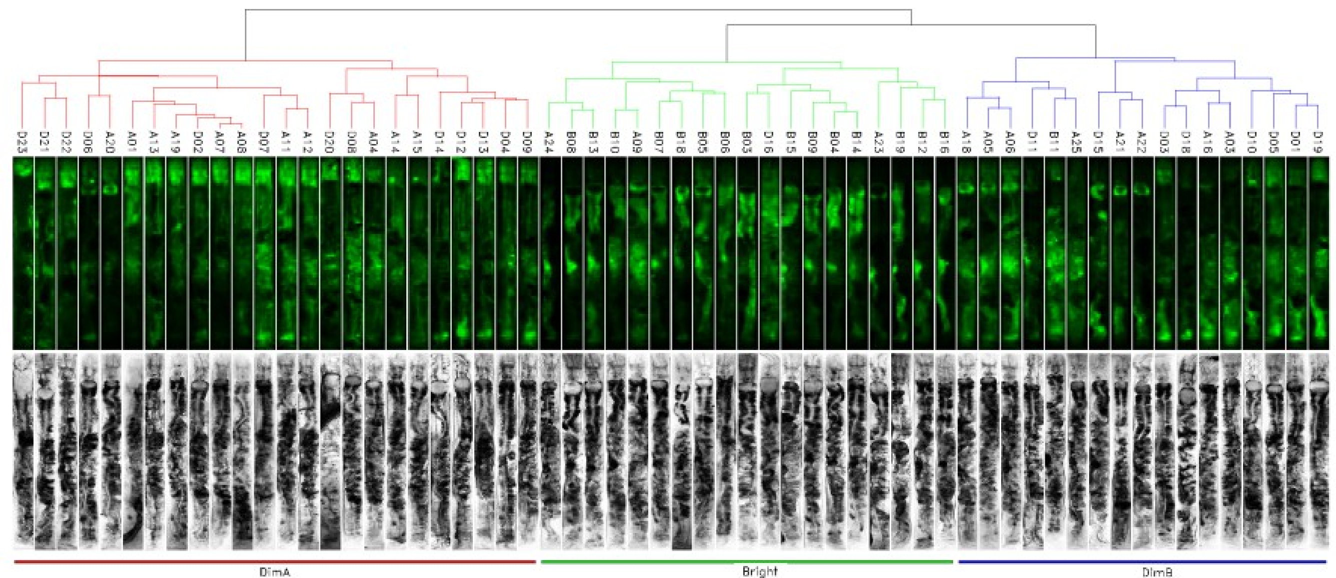
© 2012

alex@seewald.at
www.seewald.at

Questions (3)

In the result slide from *C. elegans Protein Localization (5)*, we have seen that even worms which are genetically identical (clones or „twins“) show a wide variety of different expression patterns even of a relatively ubiquitous small protein and a related large difference in expected lifetime.

Try to get into body and mind of C. elegans and discuss what might be responsible for these differences.



© 2012

SEEWALD SOLUTIONS

alex@seewald.at
www.seewald.at