technische universität
dortmund

SFB 876 Providing
Information by Resource-
Constrained Data Analysis

Universitätsklinikum Essen

UNIVERSITÄT DUISBURG ESSEN
Offen im Denken

Project C1

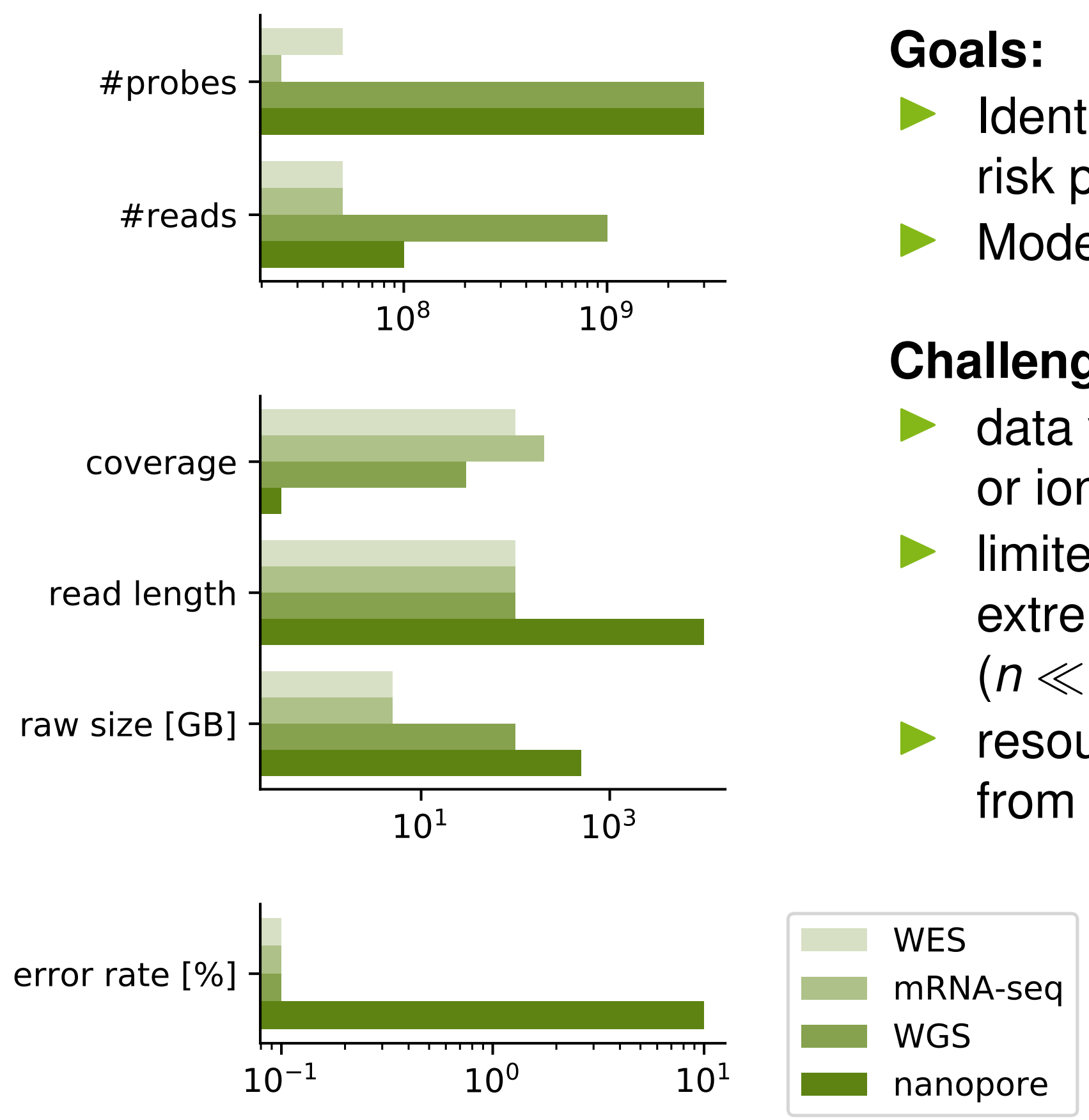
Feature selection in high dimensional data for risk prognosis in oncology

Prof. Dr. Alexander Schramm, Prof. Dr. Sven Rahmann

Problem

New Dimension of Data Volume: Whole genome & nanopore sequencing

Features are derived from molecular probes or sequences (reads).



Goals:

- Identifying molecular biomarkers for risk prognosis
- Modeling prediction functions

C4

Challenges:

- data volume (100s of GBs sequence or ion current data)
- limited number n of samples vs. an extremely high number p of features ($n \ll p$ problem)
- resource-efficient feature generation from raw data

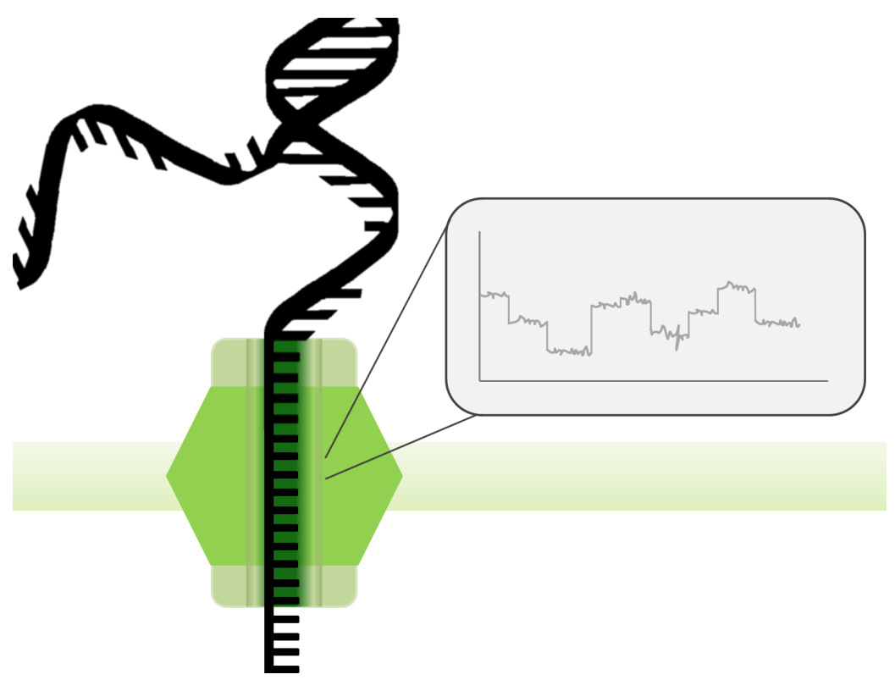
New Sequencing Technology: MinION nanopore

- Small portable device (size of a USB stick)
- Analysis on a laptop
- Large lab infrastructure no longer necessary
- Democratisation of access to sequencing
- One flow cell can generate 10–20 Gb of DNA sequence data.
- Ultra-long reads are possible (>100 Kbp)
- Disadvantage: high error rates (10–15 %)

Source: http://www.bio-itworld.com/uploadedImages/Bio-IT_World/Top_Headlines/2014/12-Dec/MinION%20close%20up.jpg

Goals:

- Real-time algorithms that convert the changes in ion currents at the nanopores immediately into the corresponding DNA sequence
- Prognostic biomarkers from liquid biopsies and nanopore sequencing
- Collaboration about detection and extraction of tumor vesicles



B2

Planned Research

Efficient Whole Genome Analysis with DNA k-mers

Use genomewide-unique k-mers ($k \in \{21, 23, 25, \dots\}$) for

- single nucleotide variant (SNV) discovery
- copy number variants (CNVs)
- structural variants (translocations, fusions)
- methylation analysis from WGBS data
- gene expression analysis from RNA-seq

Feature generation from whole genomes on a standard laptop:

- Output only unique k-mers that deviate from expected count: new k-mers, lost k-mers, surprising copy number, …
- Project deviant k-mers to biological entities (regions, genes, transcripts, pathways)
- Detect enrichment of deviant k-mers and deviant biological entities in tumour samples

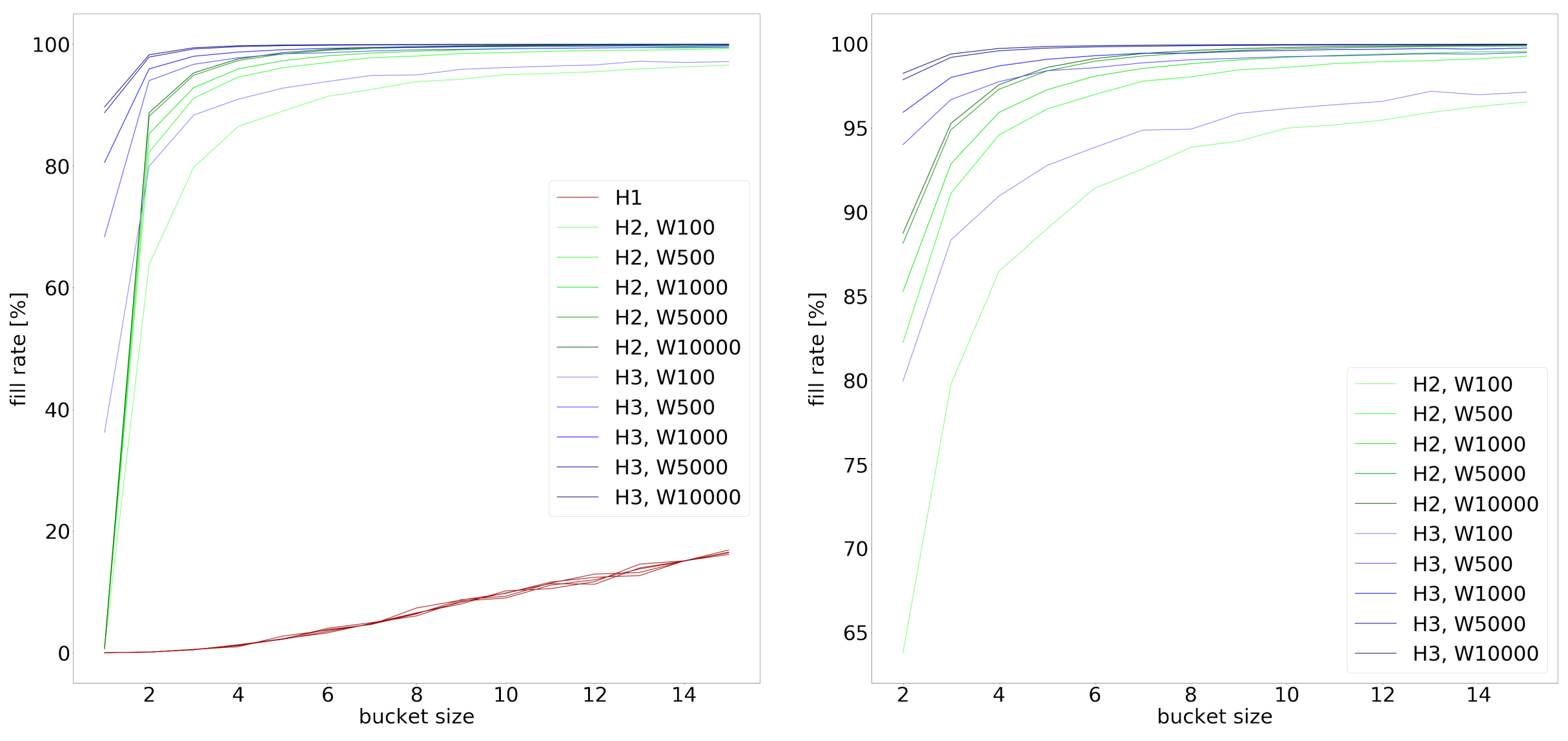
Feature reduction:

- Aggregation: Variant → Gene → Pathway
- Clustering of similar features with graph-based methods

A6

Key Data Structure: Efficient DNA k-mer Key-Value Store

Challenge: small hash table **and** fast look-ups
Speed bottleneck: cache misses during memory look-ups
New proposal: 3-way bucketed Cuckoo hashing with quotienting

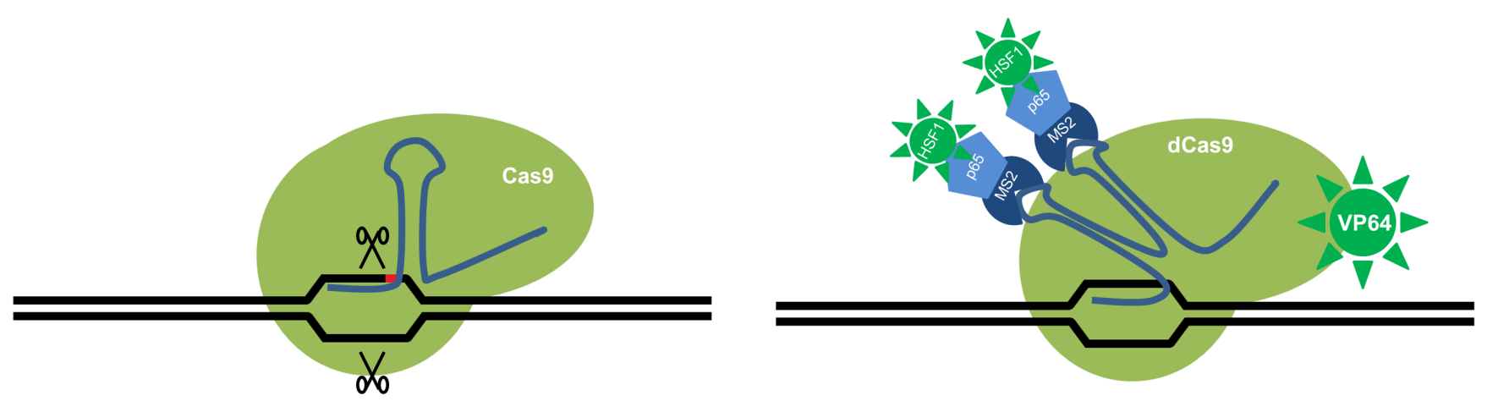


Maximal fill rates of hash table for different numbers of hash functions (H: 2 or 3), bucket sizes (x-axis, 1–15) and bounds on random walk length during insertion (W: 100, 500, 1 000, 5 000, 10 000).

Look-up needs H cache misses in the worst case.

Gene Knock Outs by CRISPR / Cas9 Genome Editing

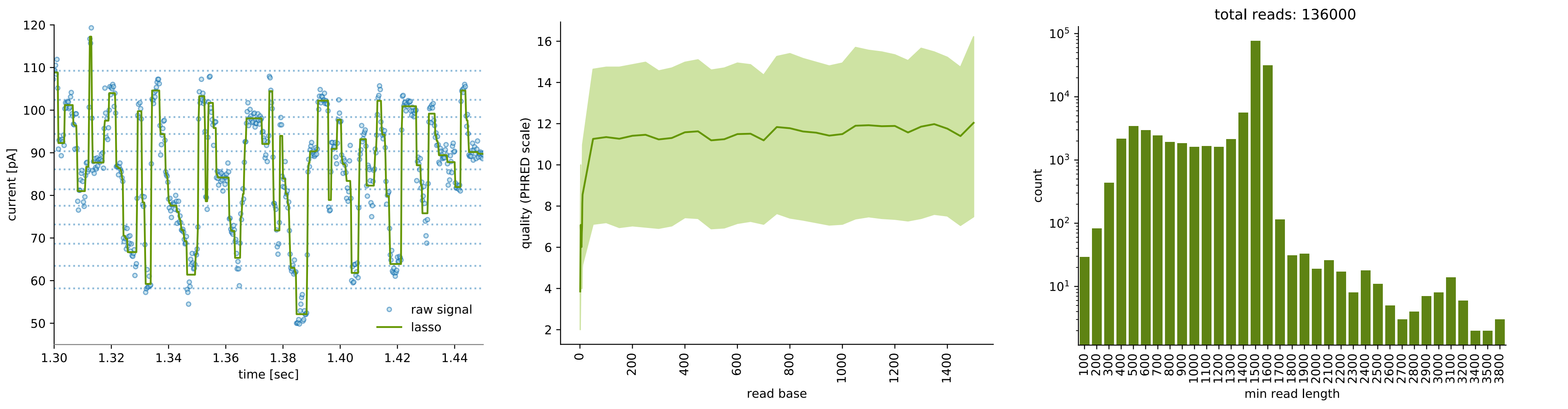
- Cultivation of human cancer cells *in vitro* (2D/3D)
- Transfection with Cas9 ribonucleoproteins or lentiviral transduction
- Double strand break generation by Cas9 to create a gene knock-out
- The inactive nuclease Cas9 complex with three activation domains activates transcription (CRISPR SAM).



- Fluorescence activated cell sorting (FACS)
- Quantitative Polymerase chain reaction (qPCR)
- MinION RNA sequencing

Analysis of Ion Current Data

Establishment of the technology and preliminary experiments on microbiomes:



Computational challenge: Lightweight conversion of ion current signal to DNA sequence

- Signal segmentation: Fused LASSO; given signal $y = (y_i)$,

$$\min \; f(x) := \sum_{i=1}^{n} (x_i - y_i)^2 + \lambda \cdot \sum_{i=1}^{n-1} |x_{i+1} - x_i|.$$

- Discretisation of signal levels; new efficient algorithms for discretised fused LASSO, where x_i must be from a finite known level set L.
- Learn mapping between k-mers of level set L to (modified) DNA sequence

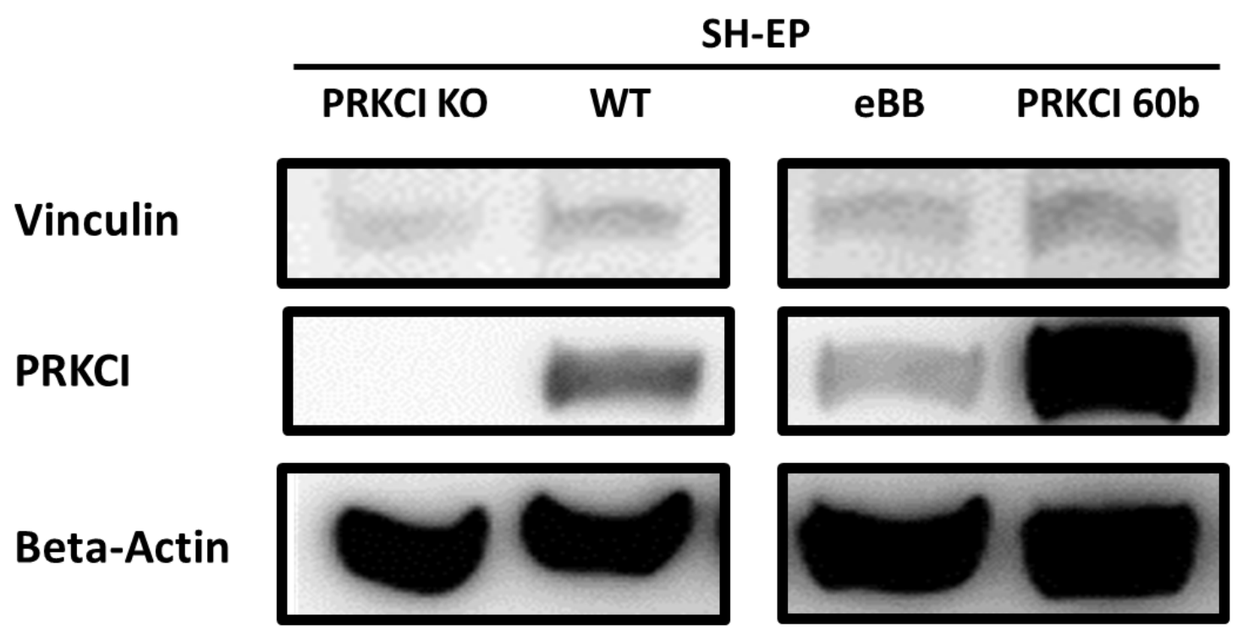
Alternative approach:

- Work with k-mers of discretised signal space L directly (richer representation)
- Discover variants as for WGS analysis in L^k-space

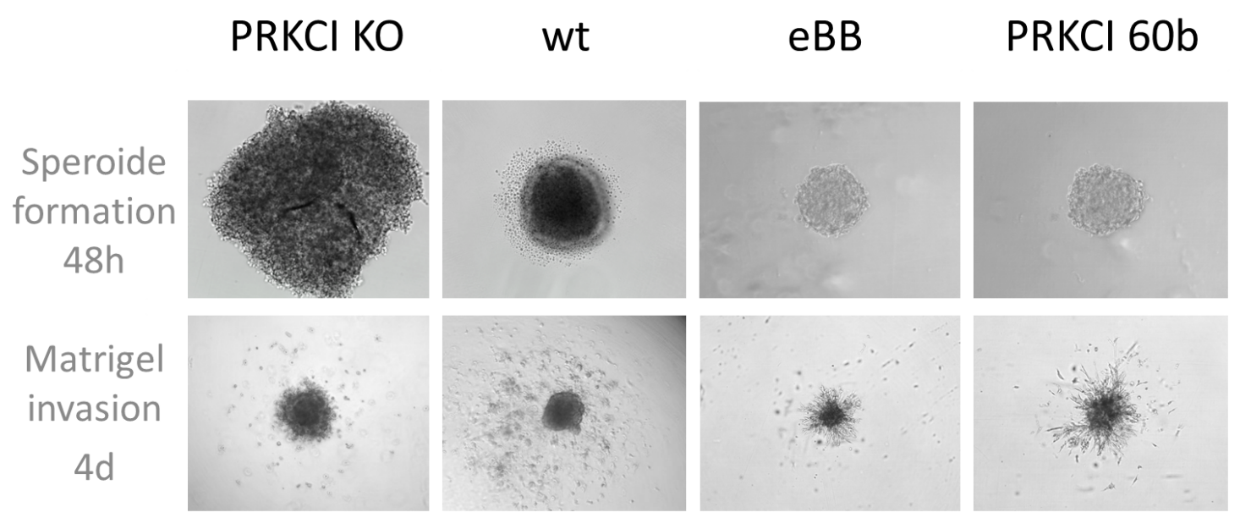
C4

Biological Target Validation

- Validation of CRISPR / Cas9 based knock-out and overexpression by Western Blot analysis



- 3D culture reveals decreased spheroid formation ability and invasiveness upon PRKCI knock-out, while over-expression of PRKCI increases the invasiveness of SH-EP cells (neuroblastoma cell line).



Molecular Oncology
Internal Medicine/Cancer Research Unit
University Hospital Essen
University of Duisburg-Essen

wtz
westdeutsches
tumorzentrum

Chair of Genome Informatics
Institute of Human Genetics
University Hospital Essen
University of Duisburg-Essen