April 19-20 2018 Paris France

RECOMB Computational Cancer Biology

ABSTRACT BOOK

Scientific Organizers:

- Valentina Boeva, Inserm & Institut Cochin, France
  Moritz Gerstung, EMBL-EBI, Hinxton, UK

Sponsor:

- Worldwide Cancer Research

Program

Thursday April 19

13:00 Welcome

13:10 – 15:00 Session 1: Genomics I

13:10 – 13:30 **Ron Zeira***: *Sorting cancer karyotypes using double-cut-and-joins, duplications and deletions*
13:30 – 13:50 **Santiago Gonzalez**: *Decrypting the evolution of somatic mutations in human cancers*
13:50 – 14:10 **Simone Zaccaria**: *Inference of allele and clone-specific copy-number aberrations in tumor samples*
14:10 – 14:30 **Linda Sunderman**: *Onctopus: Lineage-based subclonal reconstruction*
14:30 – 14:50 **Camir Ricketts***: *Meltos: Multi-sample tumor phylogeny reconstruction for structural variants*

15:00 – 16:30 Poster Session and Coffee – see p.35 for poster abstracts

16:30 – 17:30 Session 2: Immunotherapy and Other Translational Applications

16:30 – 16:50 **Pauline Depuydt**: *Genomic amplifications and distal 6q loss are novel markers for poor survival in high-risk neuroblastoma patients*
16:50 – 17:10 **Shila Ghazanfar**: *DCSR: Differential correlation across survival ranking*
17:10 – 17:30 **Hanna Najgebauer**: *CELLector: Genomics guided selection of cancer in vitro models*

17:30 – 18:30 Marta Łuksza (Keynote): Predicting cancer evolution from immune interactions

Friday April 20

09:00 – 10:20 Session 3: Single Cell Approaches

09:00 – 09:20 **Sabrina Rashid***: *Dhaka: Variational autoencoder for unmasking tumor heterogeneity from single cell genomic data*
09:20 – 09:40 **Kieran Campbell**: *Probabilistic inference of clonal gene expression through integration of RNA & DNA-seq at single-cell resolution*
09:40 – 10:00 **Yuanhua Huang**: *Cardelino: Clonal assignment of single cells with expressed mutations*
10:00 – 10:20 **Simone Ciccolella**: *Inferring cancer progression from single cell sequencing while allowing loss of mutations*

10:20 – 10:50 Coffee Break

10:50 – 12:40 Session 4: Imaging

10:50 – 11:10 **Iman Hajirasouliha**: *Classification of tumor images using deep convolutional neural networks (Highlight talk)*
11:10 – 11:30 **Yu Fu**: *Exploring the association between pathology images and genomic data in cancer*

11:30 – 12:30 Juan Caicedo (Keynote): Variant impact phenotyping using deep morphological profiling

12:30 – 14:00 Lunch (individually arranged)

14:00 – 15:20 Session 5: Functional and Systems Biology

14:00 – 14:20 **Laura Cantini**: *Stabilized independent component analysis outperforms other methods in finding reproducible signals in tumoral transcriptomes*

14:20 – 14:40 **Tiago Silva***: *ELMER 2.0: An R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles*

14:40 – 15:00 **Azim Dehghani**: *POSTIT: Multi-task learning to infer transcript isoform regulation from epi-genomics and transcriptomics data*

15:00 – 15:20 **Andre Kahles**: *Comprehensive alternative splicing analysis of 8,512 TCGA donors*

15:20 – 15:50 Coffee Break

15:50 – 18:10 Session 6: Genomics II

15:50 – 16:10 **Natalie Davidson**: *Integrative analysis of diverse transcriptomic alterations to identify cancer-relevant genes across 27 histotypes*

16:10 – 16:30 **Kjong Lehmann**: *Assessing the effect of germline and somatic mutation on gene expression changes in 1,188 human tumours*

16:30 – 16:50 **Tyler Funnell**: *Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers*

16:50 – 17:10 **Harald Vöhringer**: *TensorSignatures: a multidimensional tensor factorization framework for extraction of mutational signatures*

17:10 – 18:10 Nuria Lopez-Bigaz (Keynote): Coding and non-coding cancer mutations

18:10 Farewell

*** Talks from selected paper submissions**

Talk abstracts

Ron Zeira and Ron Shamir
School of Computer Science, Tel Aviv University, Israel

Sorting cancer karyotypes using double-cut-and- joins, duplications and deletions

Problems of genome rearrangement are central in both evolution and cancer research. Most genome rearrangement models assume that the genome contains a single copy of each gene and the only changes in the genome are structural, i.e., reordering of segments. In contrast, tumor genomes also undergo numerical changes such as deletions and duplications, and thus the number of copies of genes varies. Dealing with unequal gene content is a very challenging task, addressed by few algorithms to date. More realistic models are needed to help trace genome evolution during tumorigenesis.

Here we present a model for the evolution of genomes with multiple gene copies using the operation types double-cut-and-joins, duplications and deletions. The events supported by the model are reversals, translocations, tandem duplications, segmental deletions, and chromosomal amplifications and deletions, covering most types of structural and numerical changes observed in tumor samples. Our goal is to find a series of operations of minimum length that transform one karyotype into the other. We show that the problem is NP-hard and give an integer linear programming formulation that solves the problem exactly under some mild assumptions. We test our method on simulated genomes and on ovarian cancer genomes.

Our study advances the state of the art in two ways: It allows a broader set of operations than extant models, thus being more realistic, and it is the first study attempting to reconstruct the full sequence of structural and numerical events during cancer evolution.

Santiago Gonzalez, Moritz Gerstung and Pcawg Evolution, and Heterogeneity Working Group
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

Decrypting the evolution of somatic mutations in human cancers

Cancer is a disease characterized by an abnormal proliferation of cells. Their development is driven by the same rules that govern the evolution of living organisms: the permanent acquisition of genetic variation provided by somatic mutations and the selective pressure introduced by the environment.

At the time of the first diagnosis, when the tumour is sequenced, it has evolved for a long time accumulating large number of somatic mutations. The vast majority of them, named as passengers, have no impact on the tumour. As a consequence, it is mostly unknown when different mutations occurred and which events have determined the evolution of the tumour. However, copy number gains act as a time capsule which retains information about the state of the genome at the time the gain occurred. Using these regions, it is possible to rollback in the evolutionary history of the tumour and interrogate the events involved in the early stages.

Using this property of copy number gains, as a part of the Pan-Cancer Analysis of Whole-Genomes Consortium (PCAWG) we have analysed 2,658 different whole genome tumour samples from 39 cancer types. The results demonstrate that, in contrast to the long list of known driver genes, early stages of the tumour development are driven by a restricted set of cancer genes: 50% of the observed driver mutations in early stages can be explained by only 12 different genes. By contrast, late stages are governed by a larger set of driver genes. We also estimated when the different copy number gains and whole-genome duplications occurred. Our results suggest that whole genome duplications precede diagnosis by many years, in some cases even decades. This work is currently under revision and is available on bioRxiv (https://www.biorxiv.org/content/early/2017/08/30/161562).

In order to understand the bias of driver genes towards specific stages of tumour development and to assess the potential differential effect of mutations carried by a given gene, we have extended our analysis to 8,000 exomes from TCGA. Preliminary results suggest that, within the same gene, different positions which are recurrently mutated show similar preference to be early selected independently of how strong their recurrences are. Analysing TP53, 12 of the 14 more recurrent mutated sites show exactly the same bias towards early stages, while the remaining 2 positions, including the 5th most mutated one across all patients, have any kind of time preference. Similar results have been observed in a variety of driver genes.

These findings demonstrate how copy number gains can be used in the timing classification of mutations, providing new insights into tumour evolution, describing trajectories and unveiling the relevance of different genes across the tumour lifetime. Additionally, a more detailed analysis of the driver genes can add a new dimension to the study of variants' driver potential.

Simone Zaccaria and Benjamin Raphael
Dep. of Computer Science, Princeton University, USA

Inference of allele and clone-specific copy-number aberrations in tumor samples

Copy-Number Aberrations (CNAs) are frequent somatic mutations in tumors affecting the copy numbers of large genomic segments. CNAs are detected from differences between the observed and expected number of sequencing reads that align to a locus. Prediction of CNAs in cancer faces two main challenges. First, bulk tumor samples contain a mixture of normal cells and subpopulations of tumor cells, or clones, with different CNAs. Second, Whole-Genome Duplication (WGD) is common in solid tumors but cannot be directly identified from read counts because it doubles whole-genome content. While several methods have been developed to infer CNAs in mixed samples, most analyze single samples from a tumor and have assessed their performance by either considering real data with no ground truth or limited simulations.

In this work, we introduce a new method that infers the copy numbers and proportions of distinct clones by solving a matrix-factorization problem, and predicts the presence of WGD via a model selection step. Our method has been specifically designed to leverage the multiple samples that are routinely obtained from different regions of a primary tumor and metastasis from the same patient. We demonstrate that our method outperforms current state-of-the-art approaches, Battenberg, TITAN, THetA, and cloneHD, on 256 simulated samples, half with WGD.

Our method's advantages are particularly pronounced when multiple samples from the same tumor are jointly analyzed. Furthermore, our method shows substantially higher precision and recall in the identification of WGDs, compared to the other approaches or even a consensus of these approaches. These results suggest that our method improves copy-number inference and may be suitable for estimating the proportion of tumors with WGDs on large cohorts.

Linda K. Sundermann[1,2,*], Daniel Doerr[2], Amit G. Deshwar[3,4], Jeff Wintersinger[5], Jens Stoye[2], Quaid Morris[3,5,6,*] and Gunnar Rätsch[7,8,*]

[1]International Research Training Group GRK 1906; [1 2]Genome Informatics, Faculty of Technology and Center for Biotechnology, Bielefeld University, Germany; [3]Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada; [4]Deep Genomics Inc., Canada; [5]Department of Computer Science, University of Toronto, Canada; [6]The Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Canada; [7]Biomedical Informatics, Department of Computer Science, ETH Zürich, Switzerland; [8]Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, USA: *Corresponding authors: lsunderm@cebitec.uni-bielefeld.de, quaid.morris@utoronto.ca, gunnar.ratsch@ratschlab.org

Onctopus: Lineage-based subclonal reconstruction

Cancer cells evolve over time, leading to genetically heterogeneous cell populations. To characterize the sample, a subclonal reconstruction of these populations is essential. The subclonal reconstruction reports not only which mutations co-occur in the same subpopulations but also the proportion of cells in the sample belonging to each subpopulation, and the ancestral relationships among the subpopulations. Typical mutations, which are included in the subclonal reconstruction, are single nucleotide variants (SNVs) and copy number aberrations (CNAs).

When reconstructions are based on SNVs, the subclonal reconstruction is characterized in terms of lineages. In contrast to a subpopulation, which comprises a set of cells with the same genotype, a lineage comprises the set of all cells that descend from the same founder cell. A lineage can thus be regarded as a subtree, or clade, in a phylogenetic tree and can consist of multiple populations. In the lineage-based subclonal reconstruction, mutations are assigned to the lineage in which they first appear. The lineage frequency indicates the frequency of cells in the tumor sample that belong to this lineage. In contrast to the population-based approach which infers a complete phylogeny, ancestor-descendant relations in the lineage-based approach are only inferred if they can be observed in the data.

In contrast, subclonal reconstructions based on CNAs are characterized in terms of populations, and assign CNAs to subpopulations by reporting the absolute allele-specific copy number per population. However, if multiple CNA events affect the same genome segment, the subpopulation-based reconstruction can be ambiguous. This ambiguity can be avoided by using a lineage-based approach for CNAs, where the relative allele-specific copy number, the copy number change, is assigned to lineages.

PhyloWGS [1] and Canopy [2] are population-based subclonal reconstruction methods based on both SNVs and CNAs. Currently, they are the only two methods that infer consistent ancestral relationships between populations along the genome. Both use a Markov chain Monte Carlo (MCMC) method to sample reconstructions from the posterior distribution of their model. In contrast to Canopy, PhyloWGS does not reconstruct allele-specific CNAs but needs them as input.

Here, we present Onctopus, which is the first method that explicitly reconstructs the lineage-based subclonal composition with SNVs and CNAs where multiple CNAs can be inferred per genome segment. As input data, Onctopus receives the variant allele frequencies (VAFs) of the SNVs and the average copy number of the major

and minor alleles of genome segments. The genome is segmented in a way that consecutive regions with the same copy number are comprised in one segment. We model the VAF by assigning each SNV to one lineage, following the infinite sites assumption (ISA), while simultaneously inferring the lineage frequencies. When a CNA is present in a genome segment, the SNVs of this segment are phased relatively to this CNA. We model the observed allele-specific average copy number by assigning copy number changes to lineages and at the same time inferring the lineage frequencies. Since we permit multiple CNAs per segment, CNAs do not need to satisfy the ISA. Ancestor-descendant relationships between lineages are only inferred if they can be observed in the data. This is the case when the VAF of an SNV is influenced by a CNA or when lineages frequencies would violate the lineage divergence rule [2,3,4]. If an ancestor-descendant relation between two lineages is not observed, we mark it as potentially ambiguous. Our joint likelihood function models the VAF of the SNVs with a beta binomial distribution and the average allele-specific copy numbers with a normal distribution. We developed a linear relaxation of our model as a mixed integer linear program that can be solved with state-of-the-art solvers. After the most likely reconstruction is found by the optimization, Onctopus further disambiguates the lineage relationships. Ancestor-descendant lineage relationships can be ruled out, if an assumed ancestor-descendant relationship links SNVs of the one lineage to a copy number change in another lineage, leading to a VAF change of the SNVs.

We compared Onctopus against PhyloWGS and Canopy, following an analysis by Deshwar et al. [1] on the breast cancer data set PD4120a of Nik-Zainal et al. [3]. We ran Canopy until convergence of the MCMC sampling for three to six populations and picked the best reconstruction using Bayesian Information Criterion. For PhyloWGS, we averaged over all sampled reconstructions as recommended by the authors. Onctopus was run with four to seven lineages. In the evaluation, following Deshwar et al., we compared the inferred co-clustering of the SNVs of the chromosomes with CNAs to the co-clustering given in the analysis of Nik-Zainal et al. We used the Area under the Precision-Recall Curve (AUPRC) between the inferred and the true co-clustering matrix to compare performance. Onctopus' reconstructions outperform Canopy in three cases and are equally good in one case, and outperform PhyloWGS in all cases. A reason why Onctopus shows the best performance could be that its optimization finds the global optimum. In contrast to this, the MCMC sampling of Canopy might got stuck in a local minimum. It is possible that the MCMC sampling of PhyloWGS did not converge because the number of MCMC iterations is fixed and only one chain is used. Note that our result of PhyloWGS is different than the one in [1] as we use a different labeling to choose the SNVs for the co-clustering matrices.

As next step, we want to extend Onctopus so that it can work with multiple samples of the same cancer patient. Also, we plan to implement BIC in Onctopus, to choose the reconstruction with the best number of lineages.

[1] Amit G. Deshwar et al., "PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors." Genome biology 16.1 (2015): 35.
[2] Yuchao Jiang et al., "Assessing intra-tumor heterogeneity and tracking longitudinal and spatial clonal evolution by next-generation sequencing." (2016): 2373-2373.
[3] Serena Nik-Zainal et al., "The life history of 21 breast cancers." Cell 149.5 (2012): 994-1007.
[4] Mohammed El-Kebir et al., "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data." Bioinformatics 31.12 (2015): i62-i70.

Camir Ricketts, Daniel Seidman, Victoria Popic, Fereydoun Hormozdiari, Serafim Batzoglou, Iman Hajirasouliha
Weill Cornell Medicine, New York, USA

Meltos: Multi-sample tumor phylogeny reconstruction for structural variants

We propose Meltos, a novel computational framework to address the challenging problem of building tumor phylogeny trees using somatic structural variants (SSVs) among multiple samples.

Meltos first utilizes all possible signals for potential SV breakpoints in whole genome sequencing data, including discordant paired-end reads, split-read alignments and changes in expected depth of coverage, and proposes a probabilistic formulation for estimating variant allele fractions of SV events. Meltos then leverages the tumor phylogeny tree built on somatic single nucleotide variants (SNVs) to genotype SVs and reduce the false discovery of SV calls, using a novel optimization formulation. Our hypothesis is that small-scale somatic SVs, similar to somatic single nucleotide variations, are the result of clonal evolution in cancer samples and correspond to tumor phylogeny lineages. Thus, when we build a tumor phylogeny tree using high-quality somatic SNVs, the tree can act as a guide for calling and assigning somatic SVs.

To ensure the utility of our method on real cancer samples, we tested Meltos on multiple samples from a liposarcoma tumor. In addition to standard short-read libraries, linked-read sequencing libraries on 10X Genomics GemCode system were also provided for these samples, which gave us means for validating our results using an orthogonal technology. In addition, we tested Meltos on a multi-sample breast cancer data (Yates et al 2015), where the authors provide validated structural variation events together with deep targeted for a collection of somatic SNVs.

Pauline Depuydt[*1,2], Valentina Boeva[*3,4], Jan Koster[*5], Toby D. Hocking[6], Robrecht Cannoodt[1,2,7], Inge M. Ambros[8,9], Peter F. Ambros[8,9], Shahab Asgharzadeh[10,11], Edward F. Attiyeh[12,13,14], Valérie Combaret[15], Raffaella Defferrari[16], Matthias Fischer[17,18], Barbara Hero[19], Michael D. Hogarty[12,14], Meredith S. Irwin[20], Susan Kreissman[21], Ruth Ladenstein[8,9], Eve Lapouble[22], Geneviève Laureys[23], Wendy B. London[24], Katia Mazzocco[16], Akira Nakagawara[25], Rosa Noguera[26,27,28], Miki Ohira[29], Julie R. Park[30,31], Ulrike Pötschger[8], Jessica Theissen[17], Gian Paolo Tonini[32,33], Dominique Valteau-Couanet[34], Luigi Varesio[35], Rogier Versteeg[5], Frank Speleman[1,2], John M. Maris[12,13,14,36], Gudrun Schleiermacher[,37,38], Katleen De Preter[**,1,2]**

[1]Center for Medical Genetics, Ghent University, Ghent, Belgium; [2]Cancer Research Institute Ghent (CRIG), Ghent University, Ghent, Belgium; [3]Institut Cochin, Inserm U1016, Université Paris Descartes, Paris, France; [4]Institut Curie, Inserm U900, Mines ParisTech, PSL Research University, Paris, France; [5]Department of Oncogenomics, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands; [6]Department of Human Genetics, McGill University, Montreal, Quebec, Canada; [7]Data Mining and Modelling for Biomedicine group, VIB Center for Inflammation Research, Ghent, Belgium; [8]CCRI, Children's Cancer Research Institute, Vienna, Austria; [9]Department of Pediatrics, Medical University of Vienna, Vienna, Austria; [10]Division of Hematology/Oncology, Children's Hospital Los Angeles, USA; [11]Keck School of Medicine, University of Southern California, Los Angeles, USA; [12]Division of Oncology, Children's Hospital of Philadelphia, Philadelphia, USA; [13]Center for Childhood Cancer Research, Children's Hospital of Philadelphia, Philadelphia, USA; [14]Department of Pediatrics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA; [15]Laboratoire de Recherche Translationnelle, Centre Léon-Bérard, Lyon, France; [16]Department of Pathology, Istituto Giannina Gaslini, Genova, Italy; [17]Department of Experimental Pediatric Oncology, University Children's Hospital Cologne, Medical Faculty, University of Cologne, Cologne, Germany; [18]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany; [19]Department of Pediatric Oncology and Hematology, University Children's Hospital Cologne, Medical Faculty, University of Cologne, Cologne, Germany; [20]Division of Hematology-Oncology, Hospital for Sick Children, University of Toronto, Toronto, Canada; 21. Department of Pediatrics, Duke University School of Medicine, Durham, North Carolina, USA; 22. Genetic Somatic Unit, Institut Curie, Paris, France; 23. Department of Pediatric Hematology and Oncology, Ghent University Hospital, Ghent, Belgium; 24. Dana-Farber Children's Hospital Cancer and Blood Disorders Center, Harvard Medical School, Boston, Massachusetts, USA; 25. Saga Medical Center KOSEIKAN, Saga, Japan; 26. Pathology Department, Medical School, University of Valencia, Valencia, Spain; 27. Medical Research Foundation INCLIVA, Valencia, Spain; 28. CIBERONC, Madrid , Spain; 29. Research institute for clinical oncology Saitama Cancer Center, Saitama, Japan; 30. Seattle Children's Hospital, Seattle, Washington, USA; 31. University of Washington, Seattle, Washington, USA; 32. Laboratory of Neuroblastoma, Onco/Haematology Laboratory, University of Padua, Padova , Italy; 33. Pediatric Research Institute (IRP)-Città della Speranza, Padova, Italy; 34. Institut Gustave Roussy, Université Paris Sud, Paris, France; 35. Laboratory of Molecular Biology, Istituto Giannina Gaslini, Genova, Italy; 36. Abramson Family Cancer Research Institute, Philadelphia, PA 19104, USA; 37. Recherche Translationelle en Oncologie Pédiatrique (RTOP) and Department of Pediatric Oncology, Institut Curie, Paris, France; 38. U830, INSERM, Paris, France; * shared first authors; ** shared last authors

Genomic amplifications and distal 6q loss are novel markers for poor survival in high-risk neuroblastoma patients

Background: Neuroblastoma is characterized by substantial clinical heterogeneity. Despite intensive treatment, the survival rates of high-risk neuroblastoma patients are still disappointingly low. Somatic chromosomal copy

number aberrations have been shown to be associated with patient outcome, particularly in low- and intermediate-risk neuroblastoma patients. To improve outcome prediction in high-risk neuroblastoma, we aimed to design a prognostic classification method based on copy number aberrations.

Methods: In an international collaboration, normalized high-resolution DNA copy number data (arrayCGH and SNP arrays) from 556 high-risk neuroblastomas obtained at diagnosis were collected from nine collaborative groups and segmented using the same method. We applied logistic and Cox proportional hazard regression to identify genomic aberrations associated with poor outcome.
Results: In this study, we identified two types of copy number aberrations that are associated with extremely poor outcome. (i) Distal 6q losses were detected in 5.9% of patients and were associated with a ten-year survival probability of only 3.4%. (ii) Amplifications of regions not encompassing the MYCN locus were detected in 18% of patients and were associated with a ten-year survival probability of only 5.8%.

Conclusion: Using a unique large copy number dataset of high-risk neuroblastoma cases, we identified a small subset of high-risk neuroblastoma patients with extremely low survival probability that might be eligible for inclusion in clinical trials of new therapeutics. The amplicons may also nominate alternative treatments that target the amplified genes. Online access, analysis and visualization of the data is provided through the R2 platform and a Shiny web application.

Shila Ghazanfar[1,2,*], Dario Strbenac[2], John T. Ormerod[2,3], Jean Y. H. Yang[2,1], Ellis Patrick[2,4]

1. The Judith and David Coffey Life Lab, Charles Perkins Centre, University of Sydney, NSW 2006, Australia; 2. School of Mathematics and Statistics, University of Sydney, Sydney, NSW 2006, Australia
3 ARC Centre of Excellence for Mathematical and Statistical Frontiers (ACEMS), Richard Berry Building, The University of Melbourne, Parkville, Australia; 4. Westmead Institute for Medical Research, University of Sydney, Westmead, NSW 2145, Australia

DCSR: Differential correlation across survival ranking

Genes act as a system and not in isolation. Thus, it is important to consider coordinated changes of gene expression rather than single genes when investigating the aetiology of cancer. We have developed an approach for quantifying how changes in the association between pairs of genes may inform patient prognosis called differential correlation across survival ranking (DCSR). Modelling gene correlation across a continuous survival ranking does not require the classification of patients into ‘good’ or ‘poor’ prognosis groups and can identify differences in gene correlation across early, mid or late stages of survival outcome. When we evaluated DCSR against the typical Fisher Z-transformation test for differential correlation on real TCGA data, DCSR significantly ranked gene pairs containing known cancer genes more highly. Similar results are found with our simulation study. DCSR was applied to 13 cancers datasets in TCGA, revealing a number of distinct relationships for which survival ranking was found to be associated with a change in correlation between genes. Furthermore, we demonstrated that DCSR can be used in conjunction with network analysis techniques to extract biological meaning from multi-layered and complex data.

Hanna Najgebauer[1,2], Mi Yang[3], Hayley Francies[4], Euan A Stronach[1,5], Julio SaezRodriguez[1,2,3], Mathew J Garnett[1,4], Francesco Iorio[1,2,4]

[1]Open Targets, Wellcome Genome Campus, Hinxton, Cambridge, UK; [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK; [3]Faculty of Medicine, Joint Research Centre for Computational Biomedicine, RWTH Aachen University, Germany; [4]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK; [5]Target Sciences, GlaxoSmithKline, Stevenage, UK

CELLector: Genomics guided selection of cancer *in vitro* models

Immortalised human cancer cell lines have been the mainstay of *in vitro* oncology research for decades. Thousands of these models are commercially available and routinely used worldwide. Recent studies (Domcke et al., 2013; Jiang et al., 2016) highlighted that many of these models maybe over/under-representative of certain patient sub-populations or have acquired somatic alterations (providing selective growth advantage in culture) that are not observed in the primary disease they intend to model. This has profound implications on the quality of data and results emerging from the studies that employ these models. Therefore, there is a pressing need for robust and user-friendly computational tools that would assist researchers in the appropriate selection of the most disease-relevant *in vitro* models.

Here, we present CELLector, a computational tool implementing a new algorithm, which combines methods from micro-economy, graph theory and market basket analysis. CELLector leverages genomics data from large cohorts of primary tumours to identify relevant tumour subpopulations with corresponding genomic signatures. Subsequently, the algorithm ranks and selects cell line models based on the presence/absence of these signatures. This enables researchers to make appropriate and informed choices about model inclusion/exclusion in retrospective analyses and future studies. A key strength of CELLector is its generality; the algorithm can be applied to any disease for which *in vitro* models and matching primary/model genomics data are available. The analytical framework implemented in CELLector allows researchers to select the most clinically relevant cell line models in a genomics-guided fashion, across different cancer types, and without the need for expert knowledge about the primary disease under consideration. However, the model selection can be flexibly tailored to fit the context of a study, for example, by restricting the analysis to certain biological pathways or including/excluding a determined sub-cohort of patient genomes based on the presence/absence of a priori fixed genomic alterations (e.g. TP53 mutation).

CELLector is implemented into two distinct modules. The first module recursively identifies the most frequently occurring sets of molecular alterations (signatures) in a cohort of primary tumours, dividing it into distinct tumour subpopulations with defined molecular signatures. The second module examines the identified molecular signatures in cancer cell lines in order to identify, the most representative model(s) for each identified patient subpopulation. This approach not only maximises the covered disease heterogeneity but also enables the identification of molecular signatures underlying tumour subtypes currently lacking representative *in vitro* models, thus providing guidance for future development of new models. As a consequence, CELLector can be used to assist in the selection of *in vitro* models to maximise the translational impact of *in vitro* studies but also to systematically characterise tumour genomic subtypes of any disease cohort.

CELLector provides built-in genomics data for disease-matched primary tumours and cell lines derived from 16 cancer types, encompassing the characterisation of 4,550 tumours and ~1,300 immortalised and commercially available cancer cell lines, accounting for somatic mutations and copy number alterations for clinically relevant cancer genes (Iorio et al., 2016). The CELLector algorithm and interactive visualisation tools are implemented in an open-source R package (https://github.com/najha/CELLector) and R Shiny web application (https://ot-cellector.shinyapps.io/cellector_app/).

References

Domcke, S., Sinha, R., Levine, D.A., Sander, C., and Schultz, N. (2013). Evaluating cell lines as tumour models by comparison of genomic profiles. Nature Communications 4, 2126.

Iorio, F., Knijnenburg, Theo A., Vis, Daniel J., Bignell, Graham R., Menden, Michael P., Schubert, M., Aben, N., Gonçalves, E., Barthorpe, S., Lightfoot, H., et al. (2016). A Landscape of Pharmacogenomic Interactions in Cancer. Cell 166, 740-754.

Jiang, G., Zhang, S., Yazdanparast, A., Li, M., Pawar, A.V., Liu, Y., Inavolu, S.M., and Cheng, L. (2016). Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics 17, 525.

Marta Łuksza (KEYNOTE)
Oncological Sciences, Genetic and Genomic Sciences, Mount Sinai, New York, USA

Predicting cancer evolution from immune interactions

In recent years, novel therapies for treating cancer by means of a patient's own immune system have emerged. Checkpoint-blockade immunotherapies are designed to enable a patient's immune cells to recognize and destroy tumor cells. The process of recognition is based on specific protein binding interactions between the immune cells and cancer cells. Because these interactions depend on mutations in the cancer genome, immune recognition becomes an evolutionary problem. In this talk, I will present a new mathematical model of tumor evolution based on the fitness cost of tumor cells due to immune recognition. The model successfully predicts tumor response to checkpoint blockade immunotherapy, as shown in patient cohorts with melanoma and lung cancer. Our results highlight evolutionary similarities between cancer and viral pathogens and suggest general concepts of predictive analysis in fast-evolving systems.

BIO

Marta Łuksza is an assistant professor at the Icahn School of Medicine, Mount Sinai in New York. She completed her Ph.D in Computer Science at the Fraie Universität and the Max Planck Institute for Molecular Genetics in Berlin. Her research focuses on the evolutionary dynamics of fast-evolving systems, such us viruses, cancer and the immune system. She has developed predictive computational models for the evolution of cancer and the human influenza virus. She consults the WHO Influenza Unit in their bi-annual vaccine selection meetings.

RESEARCH INTERESTS

Marta Łuksza's research focuses on understanding biophysical mechanisms underlying the evolution of fast-adapting populations, such as pathogens and cancer cells, and at harvesting these mechanisms for predictive analysis. In particular, she is interested in how pathogen and cancer evolution is shaped by interactions with the immune system. She uses methods from machine learning, statistical physics, and information theory to address these questions. She has developed models that successfully predict the dominating influenza strains in the following season, based on genetic and biophysical properties of currently circulating viruses. These computational methods are currently used to advise the influenza vaccine selection. She also develops predictive models for cancer evolution under strong immune pressure induced by therapy. She is interested in finding genetic patterns behind immune interactions of cancer cells and in using this knowledge to predict and ultimately control progression of the disease.

Sabrina Rashid, Sohrab Shah, Ziv Bar-Joseph and Ravi Pandya
Computational Biology Department, Carnegie Mellon University, Pittsburgh, USA

Dhaka: Variational autoencoder for unmasking tumor heterogeneity from single cell genomic data

Intra-tumor heterogeneity is one of the key confounding factors in deciphering tumor evolution. Malignant cells exhibit variations in their gene expression, copy numbers, and mutation even when originating from a single progenitor cell. Single cell sequencing of tumor cells has recently emerged as a viable option for unmasking the underlying tumor heterogeneity. However, extracting features from single cell genomic data in order to infer their evolutionary trajectory remains computationally challenging due to the extremely noisy and sparse nature of the data. Here we describe 'Dhaka', a variational autoencoder method which transforms single cell genomic data to a reduced dimension feature space that is more efficient in differentiating between (hidden) tumor subpopulations. Our method is general and can be applied to several different types of genomic data including copy number variation from scDNA-Seq and gene expression from scRNA-Seq experiments. We tested the method on synthetic and 6 single cell cancer datasets where the number of cells range from 250 to 6000 for each sample. Analysis of the resulting feature space revealed subpopulations of cells and their marker genes. The features are also able to infer the lineage and/or differentiation trajectory between cells greatly improving upon prior methods suggested for feature extraction and dimensionality reduction of such data.

Kieran R Campbell[1,2,3], Alexandre Bouchard-Côté[2], Sohrab P Shah[1,3]

1. Department of Molecular Oncology, BC Cancer Agency, USA; 2. Department of Statistics, University of British Columbia, USA; 3. Department of Pathology and Laboratory Medicine, University of British Columbia, USA

Probabilistic inference of clonal gene expression through integration of RNA & DNA-seq at single-cell resolution

Human cancers form clones - sets of cells that exhibit similar mutations and genomic rearrangements. As clones evolve to resist chemotherapy understanding their molecular properties is crucial to designing effective treatments. While it is possible to measure both the DNA (that defines clonal structure) and RNA (that defines cell state) in single-cells through assays such as G&T-seq (Macaulay, 2015), such assays are time consuming and hard-to-scale. In practice, it is far more common to have large datasets where DNA and RNA are measured in separate cells through scalable technologies such as DLP sequencing (Zahn, 2017) for DNA and 10x genomics single-cell RNA-seq (Zheng, 2017). Although the destructive nature of each measurement process means the same cell will never be observed twice, if such assays are applied to the same tumor samples we expect the same clones to be present in both data views. However, it remains an open problem to link data across the expression space and genomic space that would allow for clone-specific expression estimates.

Here we present clonealign, a highly scalable statistical method to probabilistically assign each cell as measured in gene expression space (scRNA-seq) to a clone defined in copy number space (scDNA-seq) by assuming a copy-number-dependent effect on expression. We derive an expectation-maximization (EM) algorithm that parallelizes across the genes present allowing thousands of cells to be assigned to clones in minutes on commodity hardware. Through simulations we demonstrate that relatively few (<20%) genes must exhibit CNV-gene expression relationships for such assignment to be feasible and highly accurate.

We apply our method to independently generated whole genome scDNA-seq and 10x genomics scRNA-seq from a patient-derived breast cancer xenograft to characterize the gene expression of expanding clones over time. We show the method infers expected clonal proportions, and demonstrate how clonealign's clone assignments allow for prediction of held-out genes far better than could be expected at random. We also apply clonealign to a large dataset of 4000 cells from an ovarian cancer cell line, demonstrating gene expression assignment to clones defined by cutting the overall genomically inferred phylogenetic tree at different levels. Finally, we demonstrate how our framework serves as a basis for generalized multiview clustering from unpairable data sources for which we present a proof-of-concept.

Yuanhua Huang, Davis J. McCarthy, Raghd Rostom and Oliver Stegle
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

Cardelino: clonal assignment of single cells with expressed mutations

Cancer development is an evolutionary process where somatic mutations accumulate in tumour cells, resulting in a mixture of genetically distinct clones (1, 2). The genetic heterogeneity between cell sub-populations brings a major challenge to targeted cancer therapies due to different clonal susceptibility, which might causes treatment failure (3). In addition, single cell mRNA sequencing (scRNA-seq) has been used to study intratumour transcriptional variability, and very recently, it has been combined with DNA sequencing to investigate the impact of genomic alterations on gene expression, which has great potential for exploring expression signatures in inferred phylogenies (4). Since it is hard to simultaneously probe the genomic and transcriptomic information at single cell level, a two-step strategy emerges: first inferring the clonal structure from exomeseq in bulk samples, and then assigning the single cells to the inferred clones with the expressed mutations that are observed in scRNA-seq data (5, 6).

However, the clonal assignment of single cells by matching their expressed mutations to the clonal genotypes is challenging because of the following technical limitations, and the performance has not been benchmarked. First, around 90% of single nucleotide variants (SNVs) presented in bulk exome-seq are not observed in scRNA-seq, not only because of inactive transcription, but also due to the low sequencing coverages and high drop-out rates in scRNA-seq. Second, even when a heterozygous variant is expressed and sequenced, the mono-allelic expression can occur with high prevalence due to allele-specific expression. It has been reported that 76.4% heterozygous variants only express one of the two alleles in individual human primary fibroblast cells (7). This result implies a false negative error rate of 38.2%, i.e. missing the alternative reads in the expressed mutations. Therefore, it remains unclear whether, or to what degree, the single cells can be correctly assigned to clones with scRNA-seq data given such high missing rates and error rates.

In this work, we developed a statistical method, Cardelino, to infer clonal assignment of single cells and estimate the error rates of observing alternative reads in the scRNA-seq data by achieving the maximum likelihood. We also assessed the possibility of assigning single cells to known clones with scRNA-seq data by simulations. In order to precisely mimic real data, we set simulation parameters as observed or learned from 344 human fibroblast cells with 33 single-nucleotide variants (SNVs). These parameters include the clonal structure, clonal genotypes, sequencing coverages, allelic specific expression, and sequencing errors. We showed that Cardelino can accurately estimate both the false-positive error rate, which is usually small, and presumably caused by sequencing error or RNA editing, and the false-negative rate, which is usually high and probably caused by mono-allelic expression or dropout effects. Also, we can see that even with such high missing rate (86.4%) and false-negative rate (45%), 800 out 1000 cells can still be confidently assigned to specific clones (i.e., the probability gap is >0.2 between the first and the second clones), among which 76.5% cells are correctly assigned to the true clone.

In addition, we evaluated the clonal assignment systematically by varying three important parameters: the missing rate between 0.65 to 0.95, the false positive rate between 0.001 to 0.2, and the false negative rate between 0.3 to 0.65. The comparison results showed that given missing rate <0.9, false positive rate <0.1 and

false negative rate <0.55, we could achieve 70% cells being confidently assigned to clones with accuracy of 70%. Note, when we have more variants for clones, the assignability can be even more promising.

As an example, biological insights will be discussed for applying Cardelino into assigning human fibroblast cells to clones, and further investigating the transcriptional profiles for the single cells assigned to the clone that is most likely to have been induced to pluripotent stem cells. This method can also be applied to study the gene expression variability in tumour clones.

The Cardelino method is implemented in R, and freely available at https://github.com/davismcc/cardelino.

References

1. Nowell, P. C. (1976) The clonal evolution of tumor cell populations. Science, 194(4260), 23–28.
2. Schwartz, Russell and Sch¨affer, Alejandro A (2017) The evolution of tumour phylogenetics: principles and practice. Nature Reviews Genetics, 18(4), 213.
3. Greaves, M. (2015) Evolutionary determinants of cancer. Cancer discovery, 5(8), 806–820.
4. Muller, S. and Diaz, A. (2017) Single-Cell mRNA sequencing in cancer research: integrating the genomic fingerprint. Frontiers in genetics, 8, 73.
5. Tirosh, I., Venteicher, A. S., Hebert, C., Escalante, L. E., Patel, A. P., Yizhak, K., Fisher, J. M., Rodman, C., Mount, C., Filbin, M. G., et al. (2016) Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature, 539(7628), 309.
6. Kim, K.-T., Lee, H. W., Lee, H.-O., Kim, S. C., Seo, Y. J., Chung, W., Eum, H. H., Nam, D.-H., Kim, J., Joo, K. M., et al. (2015) Single-cell mRNA sequencing identifies subclonal heterogeneity in anti-cancer drug responses of lung adenocarcinoma cells. Genome biology, 16(1), 127.
7. Borel, C., Ferreira, P. G., Santoni, F., Delaneau, O., Fort, A., Popadin, K. Y., Garieri, M., Falconnet, E., Ribaux, P., Guipponi, M., et al. (2015) Biased allelic expression in human primary fibroblast single cells. The American Journal of Human Genetics, 96(1), 70–80.

Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha and Paola Bonizzoni

DISCo, University Degli Studi Milano-Bicocca, Italy

Inferring cancer progression from single cell sequencing while allowing loss of mutations

In recent years, the well-known Infinite Sites Assumption (ISA) has been a fundamental feature of computational methods devised for reconstructing tumor phylogeny trees and inferring cancer progression. However, recent studies leveraging Single Cell Sequencing (SCS) techniques showed evidence of a number of recurrence and mutational loss in several tumor samples, an observation which essentially violates a strict ISA (e.g. Kuipers *et al*., 2017, *Genome Research*). We present the SASC (Simulated Annealing Single Cell inference) tool, a new model and a robust framework based on Simulated Annealing for the inference of cancer progression from the SCS data. Our main objective is to overcome the limitations of the Infinite Sites Assumption by introducing a version of the Dollo parsimony model which indeed allows the deletion of mutations from the evolutionary history of the tumor. We demonstrate that SASC achieves high levels of accuracy when tested on both simulated and real data sets and in comparison with other available methods. The Simulated Annealing Single Cell inference tool (SASC) is open source and available at https://github.com/sciccolella/sasc **.**

Pegah Khosravi[1,2], Ehsan Kazemi [3], Marcin Imielinski[4,5,6,7], Olivier Elemento[1,2,4,7], and Iman Hajirasouliha[1,2,4,7]

1. Institute for Computational Biomeodicine, Weill Cornell Medical College, NY, USA; 2. Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA; 3. Yale Institute for Network Science, Yale University, New Haven, CT, USA; 4. Caryl and Israel Englander Institute for Precision Medicine, Weill Cornell Medical College, NY, USA; 5. Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, NY, USA; 6. The New York Genome Center, NY, USA; 7. The Meyer Cancer Center, Weill Cornell Medicine, New York, NY, USA

Classification of tumor images using deep convolutional neural networks (Highlight talk)

Automated tumor image analysis approaches have great potential to increase the precision of diagnosis in cancer patients and reduce human error.

In this project, we utilize various convolutional neural networks (CNN) architectures include basic CNN architecture, Google's Inceptions (V1 and V3), and an ensemble of Inception and ResNet with three training strategies (training the last layer of networks, training the networks from scratch, and fine-tuning the parameters).

We develop an open source pipeline for classification of different tumor images across various datasets: discrimination of (1) three cancer tissues, (2) two subtypes of lung cancer, (3) nine immunohistochemistry biomarkers, and (4) four immunohistochemistry staining scores in bladder and breast cancers. Our datasets involve 13659 whole-slide images that come from a combination of open-access histopathology images, The Stanford Tissue Microarray Database (TMAD) and The Cancer Genome Atlas (TCGA). In order to deploy the central architecture, we used a Tensorflow framework. The Python programming language version 2.7 was used for all aspects of this project. Also, TF-Slim which is a library for defining, training, and evaluating models in TensorFlow was used in this study.

On average, our pipeline achieved accuracies of 100%, 92%, 95%, and 69% for discrimination of various cancer tissues, subtypes, biomarkers, and scores, respectively. Also, our result indicated that the best performance is obtained using a pre-trained network and fine-tuning the parameters for all layers of the network.

We also compare accuracy of different strategies for training Inception-V1. In this regard, we train the model on the marker dataset of breast cancer across training the last layer, fine-tuning of the parameters for all layers, and the training of our own network from scratch. The best performance is obtained using a pre-trained network and fine-tuning the parameters for all layers of the network.

In terms of computation cost, note that we optimized our pipeline so that it can be run on CPUs. However, GPUs are indeed preferable to scale up the method to Pan-Cancer Analysis and accelerate training speed for future work. Our method yields cutting edge sensitivity on the challenging task of detecting various tumor classes in histopathology slides, reducing the false rate. Note that, our CNN_Smoothie pipeline (1) requires no prior knowledge of an image color space or any parameterizations from the users. Although the colors space for

different images have different distributions, our CNN_Smoothie method can successfully identify and register tumor variations and discriminate them consistently and robustly.

It provides pathologists or medical technicians a straightforward platform to use without requiring sophisticated computational knowledge (2). The related documentation is freely available at https://github.com/ih-lab/CNN_Smoothie.

References

(1) Deep Convolutional Neural Networks Enable Discrimination of Heterogeneous Digital Pathology Images, EBioMedicine. 27:317-328.

(2) https://news.weill.cornell.edu/news/2018/02/artificial-intelligence-aids-in-cancer-diagnosis

Yu Fu, Harald Vöhringer and Moritz Gerstung
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

Exploring the association between pathology images and genomic data in cancer

Medical imaging is an important and widely used tool in cancer diagnostics, treatment guidance and cancer evolution surveillance. Haematoxylin and eosin stained (H&E) histopathological images are collected as a clinical routine for most of the cancers for diagnosis. A large number of images have been collected by The Cancer Genome Atlas (TCGA) in conjunction with rich molecular data including genomic and transcriptomic information. The aim of this research is to uncover the association between pathology imaging, molecular and genomic data and to improve patients' clinical outcome through a more complex disease characterisation.

We collected 9,657 whole H&E slides from 8 cancer types with at least 100 normal slides as control (BRCA, COAD, KIRC, OV, LUAD, LUSC, STAD and PRAD). The whole slides are first cropped into 512 by 512 tiles, then tiles with more than 40% blank area are filtered out. A total of 4,899,918 tumour and normal tiles are then passed to a deep convolutional neural network, Inception-V3 (ref1) for classification and feature extraction from the last layer. 20% of the tiles were kept aside for an independent validation. A tumour/normal label with a likelihood and 2,048 features were computed by Inception-V3 for each tile. We obtained an average of 0.98 of training accuracy (0.97 of validation accuracy) across 8 cancer types.

A t-SNE representation (ref2) of prostate cancer samples (ref3) using 2048 features demonstrates the capacity of the extracted features to characterize morphological patterns related to high-grade or low-grade prostate cancers. Using the average of the 2048 features of all tiles per slide, we classified whole slides into different classes of Gleason scores, a grading system in prostate cancer, with a support vector machine (SVM, ref4). The classification yielded a training/validation accuracy of 1 and 0,75 respectively. Similarly, training an SVM on 3 genomic alterations that are the most prevalent in prostate cancer (ERG fusion, ETV1/4 fusion and SPOP mutation) showed accuracies ranging from [0.63-0.66].

In this study, we used an established deep learning architecture to quantify tumour morphology. These analyses showed a very high accuracy of classifying tumour/normal images and good accuracy to emulate the tumour grade. We obtained a moderate accuracy in classifying images with different genomic alterations, indicating that these mutations shape, but not exclusively dictate the resulting tumour morphology. Taken together, these data indicate that it is now feasible to quantify morphological patterns in tumour tissue slides and to correlate these data with genomic, molecular and clinical data.

References

[1] Rethinking the Inception Architecture for Computer Vision. Szegedy et al. ArXiv. 2015.
[2] Visualizing Data using t-SNE. Maaten et al. Journal of Machine Learning 2008.
[3] The Molecular Taxonomy of Primary Prostate Cancer. The Cancer Genome Atlas Research Network. Cell 2015
[4] Support Vector Machines. Marti A. Hearst. IEEE Intelligent Systems 1998.

Juan Caicedo (KEYNOTE)
Broad Institute, Boston, USA

Variant impact phenotyping using deep morphological profiling

Genome and tumor sequencing is yielding hundreds of variants associated with disease across dozens of genes, but determining causality and the functional impact of each variant is a major bottleneck. We propose morphological profiling as a rapid and inexpensive method to systematically map the impact of variants. Morphological profiling extracts single-cell measurements from microscopy images to compute signatures of treatments at high-throughput. Such signatures encode variations in cell state that can be analyzed to identify unexpected correlations between chemical and genetic perturbations. We are developing computational tools, including deep learning-based methods, to discern the functional impact of variants of unknown significance in lung cancer, and to explore how morphological profiles can complement gene expression profiles to this end.

BIO

Juan Caicedo is a postdoctoral researcher at the Broad Institute of MIT and Harvard, where he investigates the use of deep learning to analyze microscopy images. Previous to this, he studied object detection problems in large scale image collections also using deep learning, at the University of Illinois in Urbana-Champaign. Juan obtained a PhD from the National University of Colombia and completed research internships in Google Research, Microsoft Research, and Queen Mary University of London as a grad student, working in problems related to large scale image classification, image enhancement, and medical image analysis. His research interest include computer vision, machine learning and computational biology.

Laura Cantini, Ulykbek Kairov, Aurélien de Reyniès, Emmanuel Barillot, François Radvanyi and Andrei Zinovyev

Department of Bioinformatics, Biostatistics, Epidemiology and Computational Systems Biology of Cancer, Institut Curie, Paris, France

Stabilized Independent Component Analysis outperforms other methods in finding reproducible signals in tumoral transcriptomes

Motivation: Matrix factorization methods are widely exploited in order to reduce dimensionality of transcriptomic datasets to the action of few hidden factors (metagenes). Applying such methods to similar independent datasets should yield reproducible inter-series outputs, though it was never demonstrated yet.

Results: We systematically test state-of-art methods of matrix factorization on several transcriptomic datasets of the same cancer type in order to benchmark their reproducibility. Inspired by concepts of evolutionary bioinformatics, we design a new framework based on Reciprocally Best Hit (RBH) graphs. We show that a particular protocol of application of Independent Component Analysis (ICA), accompanied by a stabilisation procedure, leads to a significant increase in the inter-series output reproducibility. Moreover, we show that the signals detected through this method are systematically more interpretable than those of other state-of-art methods. We developed a user-friendly tool BIODICA for performing the Stabilized ICA-based RBH meta-analysis. We apply this methodology to the study of colorectal cancer (CRC) for which 14 independent publicly available transcriptomic datasets can be collected. The resulting RBH graph maps the landscape of interconnected factors that can be associated to biological processes or to technological artefacts. These factors can be used as clinical biomarkers or robust and tumor-type specific transcriptomic signatures of tumoral cells or tumoral microenvironment. Their intensities in different samples shed light on the mechanistic basis of CRC molecular subtyping.

Availability: The BIODICA tool is available from https://github.com/LabBandSB/BIODICA

Tiago Chedraoui Silva, Simon Coetzee, Lijing Yao, Nicole Yeager, Dennis Hazelett, Houtan Noushmehr, De-Chen Lin and Benjamin Berman
University of Sao Paolo, Brazil

ELMER 2.0: An R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles

Motivation: DNA methylation can be used to identify functional changes at transcriptional enhancers and other cis-regulatory modules (CRMs) in tumors and other primary disease tissues. Our popular R/Bioconductor package ELMER (Enhancer Linking by Methylation/Expression Relationships) provides a systematic approach that reconstructs gene regulatory networks (GRNs) by combining methylation and gene expression data derived from the same set of samples. ELMER uses methylation changes at CRMs as the central hub of these networks, using correlation analysis to associate them with both upstream master regulator (MR) transcription factors and downstream target genes. Since its initial publication in 2015, ELMER has been widely used and we identified a number of important areas for improvement.

Results: We present a major new version, (ELMER v. 2.0), which has been substantially re-written to provide greater stability, performance, and ease of use. We have allowed for greater extensibility by using standard Bioconductor data structures such as MultiAsssayExperiment, an extensible disease databank importer that supports new data sources such as the NIH Genomic Data Commons (GDC). We have added a GUI-based web interface for non-programmers, and analysis results as interactive HTML files. ELMER also now supports true genome-wide CRM identification, and functional annotation of CRMs using publicly available epigenomic data. Finally, we show that a new supervised analysis mode, which uses pre-defined sample groupings, increases statistical power and can identify additional Master Regulators, such as KLF5 in basal-like breast cancer.

Azim Dehghani Amirabad[1,2,3], Marcel H. Schulz[2,3]

1. International Max Planck Research School for Computer Science, Saarland Informatics campus, Saarbrücken, Germany; 2. Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany; 3. Excellence Cluster for Multimodal Computing and Interaction, Saarland Informatics Campus, Saarbrücken, Germany

POSTIT: Multi-task learning to infer transcript isoform regulation from epi-genomics and transcriptomics data

Inference of gene regulatory networks is a challenging problem in systems biology. Over decades, a wide spectrum of algorithms have been developed to elucidate the regulatory networks of the genome from high-throughput gene expression data. Modern RNA sequencing technologies enable the measurement isoform expression levels beyond classical gene expression. In this study, we challenge the traditional view of the virtual gene-level analysis by demonstrating that using gene expression data can lead to sub-optimal inference of the regulatory networks. To address this, we introduce Transcript Isoform Regula- tory Networks (TIRNs). We show that TIRNs provide an enhanced picture of genomic regulation, which overcome limitations of gene-level based approaches.

We introduce a multi-task regression framework, which integrates epi-genomic information (e.g. DNase1-seq or ATAC-seq) and transcriptomic (RNA-seq and small RNA) data into the model. We use a structured sparsity inducing penalty to incorporate alternative promoter and 3'-UTR structure of the transcript isoforms into a multi-task regression objective function that can accurately handle the complexity of transcript annotations. We present a novel and efficient proximal gradient descent based algorithm to solve the multi-tasking objective for large instances, such as the complete human transcriptome. Our algorithm, called POSTIT, enables for the first time, the inference of comprehensive TIRNs, mediated by transcription factors (TFs), RNA binding proteins (RBPs) and micro RNAs.

Evaluations on simulated data and TCGA cancer data show that POSTIT has superior performances over single task learning methods. Analysis of model performance in transcript expression prediction, highlights the contribution of regulatory elements in 3'-UTRs and alternative promoters of transcripts. More- over, it pinpoints important TFs and RBPs contributing to transcript isoform switch events in cancer. Pan-cancer analyses allowed us to infer cancer-recurrent transcript isoform regulatory modules. We provide an R and MATLAB implementation of POSTIT.

Andre Kahles, Kjong-Van Lehmann, Nora C Toussaint, Matthias Hüser, Stefan G Stark, Timo Sachsenberg, Oliver Stegle, Oliver Kohlbacher, Chris Sander and Gunnar Rätsch
ETH Zurich, Department of Computer Science, Switzerland

Comprehensive alternative splicing analysis of 8,512 TCGA donors

We present a comprehensive analysis of alternative splicing across 32 TCGA cancer types of 8,512 patients. We detect alternative splicing events (ASE) and tumor variants by reanalyzing TCGA RNA and whole exome sequencing data. Many tumors have thousands of ASE not detectable in TCGA normal and GTEx samples. Overall, tumors have ≈20% more ASE than normal samples. Association analysis of somatic variants with ASE confirmed known trans associations with variants in SF3B1 and U2AF1 and identified 7 additional trans-acting variants (including in IDH1, ANAPC1). On average, we identified ≈740 novel introns ("neointrons") in tumor samples not found in normal samples. From protein mass spectra available for TCGA breast, ovarian and colorectal tumor samples, we confirmed on average 16 peptides per sample corresponding to ASE and predicted as MHC-I binders (compared to only ≈1 SNV-associated peptide per sample). Tumor-specific splicing presents a large new class of splicing-associated potential neoantigens that may affect the immune response and could be potentially exploited in immunotherapy, e.g., in personalized tumor vaccines.

Natalie Davidson, Alvis Brazma, Angela Brooks, Claudia Calabrese, Nuno A. Fonseca, Jonathan Goeke, Andre Kahles, Kjong-Van Lehmann, Gunnar Ratsch, Roland Schwarz, Zemin Zhang, Fenglin Lui and Deniz Demircioglu

Memorial Sloan Kettering Cancer Center, New York, USA

Integrative analysis of diverse transcriptomic alterations to identify cancer-relevant genes across 27 histotypes

Background: Previous multi-cancer genomic studies have focused on the analysis of somatic mutations as the driver of phenotypic changes. Here, we propose and apply a novel method to integrate a wide variety of transcriptomic aberrations in combination with DNA-level changes to redefine the concept of driver events and account for the role of the transcriptome in tumorigenesis.

We present a novel analysis that 1) identifies cancer driver genes through a recurrence analysis over diverse types of transcriptomic alterations 2) identifies frequent and heterogeneous transcriptomic alteration signatures in 1,188 samples across 27 histotypes as part of the PanCancer Analysis of Whole Genomes (PCAWG) of the International Cancer Genome Consortium (ICGC). We integrated the following alteration types: expression outliers, alternative splicing outliers, gene fusions, alternative promoters, non-synonymous variants, RNA-editing, and allele-specific expression.

Results: To identify the cancer relevant genes, we created a new method for performing recurrence analysis on transcriptomic features. Our method has three main strengths: flexibility to handle any number or type of alteration, sensitivity to different frequencies of alterations, and ability to prioritize genes with heterogeneous alterations. The method has four main steps: 1) binarize each alteration type by identifying rare/outlier events; 2) sum alteration events over samples for each gene; 3) transform counts to ranks for alteration comparability; 4) combine ranks across alterations to create ranking scores. We performed >1M permutations to identify a cut-off for informative scores; our analysis yielded 1,012 genes with an empirical p-value <0.05. These genes show a 2.82-fold enrichment for cancer census genes (CGC) and driver genes (provided by ICGC Driver Working group;PCAWG-9) with a p-value of 5x10-26 (hypergeometric test), signifying that our analysis is identifying cancer-relevant genes.

Among the top 5% of our ranked genes is CDK12, which is impacted by multiple, but non-overlapping, types of alterations within a protein kinase domain associated with dysregulation of DNA repair in cancer. In our cohort, we find 87 samples that have an alteration in this domain, with 64 (74%) samples having only a RNA alteration in the domain. The most frequent alteration is an alternative promoter event that leads to a truncated transcript of CDK12, removing a majority of the kinase domain. Fusion and splice events lead to additional disruptions of the same domain. CDK12 exemplifies the value of integrating RNA and DNA alterations.

To understand the functional impact of these alterations, we compare alteration patterns across cancer types and cancer-specific pathways. We notice that alteration patterns vary between the cancer types: Chromophobe renal cell carcinoma (Kidney-ChRCC) in comparison with Skin-Melanoma has significantly different numbers of non-synonymous variants (t-test; p-adj.: 1.42x10-5), copy-number alterations (p-adj.: 6.70x10-4), fusions (p-adj.: 1.56x10-4), and splice outliers (p-adj.: 7.05x10-10). In contrast, similar cancers like Kidney-RCC and Kidney-ChRCC only differ in the amount of non-synonymous variants (t-test; p-adj.: 5.50x10-25). Comparing across

cancer-specific pathways, we find that the TOR and metabolism pathways are more impacted by RNA alterations. We also find that of the 578 samples with an altered p53 pathway, typically associated with high non-synonymous variants, 131 (22.7%) of them carried only RNA alterations. This is further evidence that neglecting transcriptomic alterations could underestimate the degree of cancer pathway disruption.

Conclusions: Through our joint analysis, we integrated both DNA and RNA aberrations in a recurrence analysis that yielded a list of promising genes highly enriched for known cancer driver genes. Furthermore, we identified associations of transcriptomic alterations to cancer type and DNA-level aberrations, helping to broadly explain the influence of the transcriptome on various cancer-related processes.

Lehmann K[2,3,4,*,$], Calabrese C[1,*,$], Urban L[1,5,*,$], Liu F[7,$], Erkek S[5], Fonseca NA[1], Kahles A[2,3], Kilpinen H[10,11], Markowski J[6], PCAWG Group 3, Waszak SM[5], Korbel JO[5], Zhang Z[7], Brazma A[1,#], Rätsch G[2,3,4,8,9,#], Schwarz RF[1,6,#], Stegle O[1,5,#]

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK; 2. Department of Computer Science, ETH Zürich, Switzerland; 3. Memorial Sloan Kettering Cancer Center, New York, USA; 4. University Hospital Zurich, Switzerland; 5. European Molecular Biology Laboratory, Heidelberg, Germany; 6. Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Berlin, Germany; 7. Beijing Advanced Innovation Center for Genomics and College of Life Sciences, Peking University, China; 8. Department of Biology, ETH Zürich, Switzerland; 9. Weill Cornell Medical College, New York, USA; 10. UCL Great Ormond Street Institute of Child Health, University College London, UK; 11. Wellcome Trust Sanger Institute, Hinxton, UK; $ First authors; # Last authors in alphabetical order.

Assessing the effect of germline and somatic mutation on gene expression changes in 1,188 human tumours

To systematically assess regulatory effects of germline and somatic variants on a genome-wide scale, we analyze matched whole-genome sequencing and RNA-seq data of 1,188 cancers from the TCGA/ICGC PanCancer Analysis of Whole Genomes (PCAWG) across 27 cancer types. We map expression quantitative trait loci (eQTL) of germline variants identifying 2,509 genes with a germline effect (FDR < 5%), including prominent cancer genes. A systematic comparison for replication of this eQTL map to regulatory variants in matched normal tissues from GTEx, resulted in 426 cancer-specific eQTLs that cannot be recapitulated in GTEx while all other eQTL did replicate in at least one GTEx tissue. 98 of these genes showed cancer specific upregulation and ectopic expression among which we find genes like SLAMF9 with known roles in immune response and cancer.

Using this germline map, we quantify the amount of the regulatory effect of germline mutations in different genomic regions and contrast them to the effect of somatic mutational burden. We find that SCNA is a major driver of expression variation (27.3% on average) followed by somatic mutation in flanking regions of the individual genes and nearby germline variants.

We assessed multiple strategies for association between somatic mutational burden and gene expression to understand the effect of mutational load between lymphomas and carcinomas as well as the effect of purity on detection power. This helped us to identify 649 somatic eQTL (FDR<5%) among which several genes have known roles in the pathogenesis of specific cancers (e.g., CDK12, IRF4, etc.). Tests for enrichment of regulatory annotations identified several epigenetic annotations including poised promoters and enhancers but no significant enrichment of TFBS.

This study enabled us for the first time to assess the transcriptional landscape of 1,188 cancer patients with both whole genome and gene expression data, allowing us to gain insights into the germline and somatic mechanisms in non-coding regions driving changes in gene expression. In addition to confirming known regulatory effects, we identified novel associations between somatic variation and expression, in particular in distal regulatory

elements. This work represents the first large-scale assessment of the effects of both germline and somatic genetic variation on gene expression in cancer and creates a valuable resource cataloguing these effects.

Tyler Funnell, Allen Zhang, Yu-Jia Shiah, Diljot Grewal, Robert Lesurf, Steven McKinney, Ali Bashashati, Yi Kan Wang, Paul Boutros and Sohrab Shah
BC Cancer Research Centre, Canada

Integrated single-nucleotide and structural variation signatures of DNA-repair deficient human cancers

Patterns of mutation in cancer genomes reflect both endogenous and exogenous mutagenic processes, allowing inference of causative mechanisms, prognostic associations, and clinically actionable vulnerabilities in tumors. Many mutational processes leave distinct genomic "footprints", measurable via nucleotide substitution patterns, localised mutation densities, and patterns of structural variation. As such, each mutagenic source (whether exogenous or endogenous) changes DNA in a characteristic manner, at genomic locations with preferred chemical and structural characteristics.

Cancer cells can also acquire an endogenous mutator phenotype, accumulating large numbers of mutations due to DNA repair deficiencies. Defective DNA repair processes induce both point mutations and structural variations, and include several mechanistic classes such as mismatch repair deficiency, homologous recombination deficiency, microhomology mediated end-joining, and breakage fusion bridge processes. Defective DNA repair has been exploited in therapeutic regimes, including immune checkpoint blockade for mismatch repair deficiency, and synthetic lethal approaches for homologous recombination deficiency, underscoring their clinical importance.

Both point mutation signatures and structural variation signatures have been studied extensively as independent features of cancer genomes, mostly through non-negative matrix factorization (NMF) approaches. As increasing numbers of whole genomes are generated from tumors in international consortia and focused investigator research, the need for robust signature inference methods is acute. Additional computational methods have been proposed, however no approaches jointly infer signatures from both point mutation and structural variations. We contend that systematic, integrative analysis of point mutation and structural variation processes enhances the ability to exploit signatures for subgroup discovery, prognostic and therapeutic stratification, clinical prediction, and driver gene association.

Topic models, a popular and effective class of methods for natural language document analysis, are well suited to the task of mutation signature inference. In this paper we introduce the correlated topic model, which incorporates signature activity correlation, and a multi-modal correlated topic model (MMCTM) which jointly infers signatures from multiple mutation types, such as SNVs and SVs. Signature activities can be correlated among some groups of patients, motivating the use of this class of methods. For example, homologous recombination deficiency induces patterns of both SNVs and SVs in breast and high grade serous ovarian cancers. We show how integrating SNV and SV count distributions improves inference of signatures relative to NMF and standard topic modeling methods.

Motivated by the need to better understand mutation signatures in the context of DNA repair deficiency, we applied the MMCTM to SNV and SV somatic mutations derived from publically available breast and ovarian whole genomes, performing joint statistical inference of signatures. Our results reveal correlated topic models as an important analytic advance over standard approaches. Rigorous benchmarking over mutation signatures inferred from previously published breast cancer mutation corpora was used to establish metrics for comparison. In addition, we report a novel, prognostically relevant ovarian cancer patient group using MMCTM-derived SNV and SV signature profiles. In aggregate, our study reveals the importance of simultaneously considering multiple classes of genomic disruption as a route to expanding mutation signature discovery, and their downstream impact on novel stratification across human cancers.

Harald Vöhringer and Moritz Gerstung
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

TensorSignatures: a multidimensional tensor factorization framework for extraction of mutational signatures

Cancer arises through the accumulation of mutations caused by multiple processes that each leaving behind distinct patterns of mutations on the DNA. A number of studies have analysed cancer genomes to extract such mutational signatures using non-negative matrix factorisation (NMF) over catalogues of single nucleotide variants (SNVs).

However, many mutational processes also generate characteristic multi nucleotide variants (MNVs), insertions and deletions (indels), and structural variants (SVs). Moreover, it has been reported that mutagenesis also depends on the transcriptional orientation and the direction of replication origin, and sometimes manifests as local hypermutation (kataegis). These phenomena add additional features and dimensions to the mutation data, which make the resulting multidimensional tensor data unamenable to conventional matrix factorisation methods. Finally, there is evidence that mutation counts can vary substantially and thus require overdispersed probabilistic models to avoid overclustering.

We developed TensorSignatures - a multidimensional tensor factorisation framework incorporating the aforementioned phenomena for a more comprehensive and robust extraction of mutational signatures.

We have tested the algorithm using simulations and on a dataset comprising 2778 whole genomes from the ICCG PCAWG consortium spanning 39 cancer types.

The resulting signatures resemble known processes such as BRCA deficiency, APOBEC hypermutation, or mismatch repair deficiency, avoid oversegmentation and add additional biological detail. For example, the resulting BRCA signature contains large deletions and tandem duplications and is highly discriminative from other processes. Also, we quantify wide-spread transcriptional biases including UV, tobacco smoking signatures and in liver cancer indicating transcription coupled repair and damage processes. Replication biases are most pronounced in polymerase epsilon deficiency, mismatch repair deficiency and APOBEC signatures, indicating that these operate during DNA replication. Locally clustered mutations stem mostly from mutational processes involving APOBEC and polymerase eta associated hypermutation introducing about 10% clustered mutations.

In summary, we present a highly robust and flexible tensor factorisation algorithm for mutational signature analysis. The resulting signatures are a comprehensive and accurate reflection of the underlying mutational processes. Lastly, the general framework can be easily extended to incorporate additional genomic features in future studies.

Nuria Lopez-Bigaz (KEYNOTE)
ICREA-Institue for Research in Biomedicine (IRB Barcelona), Barcelona, Spain

Coding and non-coding cancer mutations

Somatic mutations are the driving force of cancer genome evolution. The rate of somatic mutations appears to be greatly variable across the genome due to variations in chromatin organization, DNA accessibility and replication timing. However, other variables that may influence the mutation rate locally are unknown. I will discuss recent findings from our lab on how DNA-binding proteins and differences in exons and introns influence mutation rate. These finding have important implications for our understanding of mutational and DNA repair processes and in the identification of cancer driver mutations. Given the evolutionary principles of cancer, one effective way to identify genomic elements involved in cancer is by tracing the signals left by the positive selection of driver mutations across tumours. We analyze thousands of tumor genomes to identify driver mutations in coding and non-coding regions of the genome.

BIO
Nuria Lopez-Bigas is a biologist with a PhD in molecular genetics. She transitioned into bioinformatics during his postdoc at the European Bioinformatics Institute (EBI). Since 2006 she leads a research group in Barcelona focused on the study of cancer from a genomics perspective. She is particularly interested in the identification of cancer driver mutations, genes and pathways across tumor types and in the study of somatic mutational processes.

Among the most important achievements obtained by Lopez-Bigas' lab are the development of pioneer methods to identify cancer driver genes (Oncodrive methods) and the creation of tools for the analyses of cancer genomes: IntOGen (http://www.intogen.org) and Cancer Genome Interpreter (http://www.cancergenomeinterpreter.org). Her lab has recently discovered that DNA-bound proteins interfere with the nucleotide excision repair machinery, leading to increased rate of DNA mutations at the protein binding sites. They have also demonstrated that exons have a reduced mutation rate due to higher mismatch-repair activity in exonic than in intronic regions.

Poster abstracts

CCB1. Joao M. Alves and David Posada
Evolutionary and Biomedical Genomics, University of Vigo, Spain

Sensitivity to sequencing depth in single-cell cancer genomics

Introduction: Querying cancer genomes at single-cell resolution is expected to provide a powerful framework to understand in detail the dynamics of cancer evolution. However, given the high costs currently associated with single-cell sequencing (SC-Seq), together with the inevitable technical noise arising from single-cell genome amplification, cost-effective strategies that maximize the quality of single-cell data are critically needed. Taking advantage of publicly available cancer datasets, we studied the impact of sequencing depth and sampling effort towards single-cell variant detection, including structural and driver mutations, genotyping accuracy, clonal inference and phylogenetic reconstruction, using recent tools specifically designed for single-cell data.

Methods: Five single-cell whole-genome and whole-exome cancer datasets from four distinct SC-Seq studies were retrieved from the Sequence Read Archive, including four single-cell genomes from a breast cancer patient (we call this dataset W4 to indicate the authors and the number of cells), eight single-cell exomes from circulating tumor cells from one lung adenocarcinoma patient (N8), 25 single-cell exomes derived from a kidney tumor patient (X25), 55 single-cell exomes from a breast cancer patient (W55), and 65 single-cell exomes from a single JAK-2 negative neoplasm myeloproliferative patient (H65). Single-cell data were independently downscaled to 25, 10, 5, and 1x. For each depth level, ten technical replicates were generated for statistical validation, resulting in a total of 6280 single-cell BAM files.

Results: Altogether, our results suggest that, even though sequencing depth does indeed contribute to a better refinement of somatic variant characterization from tumor single cells, sample size plays a more determinant role for a reliable assessment of the general patterns of somatic variation in cancer genomes. For relatively large sample sizes (e.g., $\geq$25 samples), sequencing single cells at modest depths (i.e., 5x) enables a similar description of somatic variation, clonal composition, and evolutionary history compared to sequencing depths one order of magnitude higher.

Conclusions: Here, we demonstrate that sequencing many individual tumor cells at a modest depth represents an effective alternative to explore the mutational landscape and clonal evolutionary patterns of cancer genomes, without the excessively high costs associated with high-coverage genome sequencing.

CCB2. Gurkan Bebek and Arda Durmaz
Centre for Proteomics and Bioinformatics, Case Western Reserve University, Cleveland, USA

Frequent subgraph mining of dysregulated pathways predict effective cancer therapeutics

Motivation: Cancers are comprised of a distinct set of mutations and effective cancer therapeutics target these alterations to treat the disease. Large-scale genomics studies have generated comprehensive molecular

characterization of numerous cancer cell lines to better understand dysregulation caused by molecular events leading to disease. We hypothesize that frequent graph mining of pathways to gather pathways functionally relevant to cancer cell lines can characterize altered states and provide opportunities for identifying effective drugs leading to developing personalized therapies.

Results: We describe an integrative omics approach based on frequent subgraph mining (FSM) where Protein-Protein Interaction (PPI) networks and gene expression data are used together to infer dysregulation in samples. We identified altered pathways in cancer cell lines previously analyzed. We grouped cell lines through these altered pathways in an unsupervised fashion. Cancer cell line responses to therapeutic compounds were mapped to these group of cell lines and we investigated how the average similarity within drug groups affect cell lines response to treatment.

Conclusions: The models and predictors we report could serve as a means to provide opportunities for improved treatment options and personalized interventions. Our proposed novel graph mining approach is able to integrate PPI networks with gene expression in a biologically sound approach.

CCB3. Mariana Buongermino Pereira[1,2], Erik Delsing Malmberg[3], Anna Rehammar[1], Jonas Abrahamsson[4], Tore Samuelsson[5], Sara Ståhlman[6], Julia Asp[3,6], Lars Palmqvist[3,6], Erik Kristiansson[1], Linda Fogelstrand[3,6]

1. Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden; 2. Research Department of Oncology, Cancer Institute, Faculty of Medical Sciences, University College London, UK; 3. Department of Clinical Chemistry and Transfusion Medicine, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden; 4. Department of Pediatrics, Institute of Clinical Sciences, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden; 5. Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy at University of Gothenburg, Gothenburg, Sweden; 6. Department of Clinical Chemistry, Sahlgrenska University Hospital, Gothenburg, Sweden

Accurate and sensitive analysis of minimal residual disease in acute myeloid leukemia using deep sequencing of single nucleotide variations

An accurate, sensitive and broadly applicable method for minimal residual disease (MRD) analysis in acute myeloid leukemia (AML) would enable a better risk stratification and disease monitoring for treatment decisions. Currently, MRD monitoring in AML is supported by multiparameter flow cytometry (MFC), a technique that is broadly applicable, but has a low sensitivity as 20%-40% of MFC-MRD negative patients relapse. Alternatively, somatic mutations found in the leukemic cells can be used as MRD markers. We recently showed that such leukemia-specific mutations can be identified with exome sequencing at diagnosis and assessed during follow-up at low frequencies using targeted deep sequencing. Here, we applied a statistical model to correct for position-specific sequencing errors in targeted-deep sequencing by using position-specific reference samples. The variant allele was described using a binomial model where an additive variance component where introduced to characterise the between-sample variability. The model was trained using 18 control samples from healthy volunteers. The model achieved a limit of detection for single nucleotide variations of variant allele frequency (VAF) 0.02%. When the assay with this model was evaluated in AML patient samples it showed good

concordance with MFC-MRD but higher sensitivity. The method was linear in MRD range (0.03-1%) with good precision and low bias. In conclusion, the application of this statistical model to targeted deep sequencing enables highly sensitive and reliable detection of low allelic frequencies without the need for unique molecular identifiers. The introduction of this method in patient care will allow for highly sensitive disease monitoring in virtually every patient with AML.

CCB4. Kieran R Campbell[1,2,3], Alexandre Bouchard-Côté[2], Sohrab P Shah[1,3]

1. Department of Molecular Oncology, BC Cancer Agency, Canada; 2. Department of Statistics, University of British Columbia, Canada; 3. Department of Pathology and Laboratory Medicine, University of British Columbia, Canada

Probabilistic inference of clonal gene expression through integration of RNA & DNA-seq at single-cell resolution

See abstract at p.15.

CCB5. Laura Cantini, Ulykbek Kairov, Aurélien de Reyniès, Emmanuel Barillot, François Radvanyi and Andrei Zinovyev

Department of Bioinformatics, Biostatistics, Epidemiology and Computational Systems Biology of Cancer, Institut Curie, Paris. France

Stabilized Independent Component Analysis outperforms other methods in finding reproducible signals in tumoral transcriptomes

See abstract at p.23.

CCB6. Simone Ciccolella, Mauricio Soto Gomez, Murray Patterson, Gianluca Della Vedova, Iman Hajirasouliha and Paola Bonizzoni

DISCo, University degli Studi Milano-Bicocca, Italy

Inferring cancer progression from single cell sequencing while allowing loss of mutations

See abstract at p.18.

CCB7. Rui Costa and Moritz Gerstung
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

A hierarchical Bayesian multistage Cox model for high-dimensional disease progression analysis

Survival analysis is instrumental in establishing the risk factors for cancer. As genomic-clinical data sets become larger and more detailed, the demand grows for statistical methods that can deal with large amounts of information and deliver more insightful analyses. In particular, methods are sought that can help identify driver genes, deliver individual prognoses, and eventually support personalised clinical decisions.

Standard approaches in survival analysis are faced with essentially two challenges. First, detailed clinical histories require methods that are flexible enough to analyse complex disease progression data, rather than just single-event (illness or death) or repeated-event observations. In oncology, a typical example is that of a patient with acute myeloid leukemia who has progressed through different types of myeloproliferative neoplasms. Recent R packages, such as MSM and MSTATE, achieve the required flexibility in the context of the multistate Cox model. Second, the incorporation of higher-dimensional covariate data, such as mutation and gene expression data, is not possible without some degree of model regularisation, out-of-sample validation, or other procedure for overfitting reduction. For example, regularised estimation of the (single-event) Cox model can be obtained with the R package COXME, by defining random covariate coefficients.

Computer packages that are designed for arbitrary multistate disease histories and also contain a regularisation feature are scarce. To the best of our knowledge, the only such packages in the CRAN repository are penMSM and gamboostMSM, which perform estimation of the multistate Cox model using fusion lasso and gradient boosting respectively. Our own contribution to fill this gap is also an implementation of the multistate Cox model. It makes however use of different regularisation procedures: overfitting can be reduced either by defining a hierarchical Bayesian model or by running the complementary pairs stability selection algorithm. In the case of the hierarchical Bayesian model, it is assumed that the parameters follow normal distributions. Multiple groups of parameters can be defined, with parameters within each group sharing the same mean and variance. The group means and the variance components of the model are efficiently estimated by an EM-type algorithm.

In contrast to penMSM and gamboostMSM, our method can estimate, for any time since the onset of the disease, the probability that a particular patient is in a given state, and it can build sediment plots of these occupation probabilities, as multistate analogues of the classical survival curves. In addition, it can go beyond the traditional markovian Cox model, by making the exit transition rates from any given state depend on the time spent in that state. The key functions of our method are illustrated on a genomic-clinical data set of 2040 patients with myeloproliferative neoplasms.

CCB8. Tiffany Delhomme, Patrice H. Avogbe, Aurelie Gabriel, Nicolas Alcala, Noemie Leblay, Catherine Voegele, Priscilia Chopard, Amelie Chabrier, Behnoush Abedi-Ardekani, Valerie Gaborieau, Graham Byrnes, Florence Le Calvez-Kelm, Ghislaine Scelo, Lynnette Fernandez-Cuesta, Paul Brennan, James D. McKay and Matthieu Foll

International Agency for Research on Cancer (IARC-WHO), Lyon, France

Needlestack: an ultra-sensitive variant caller for multi-sample deep next generation sequencing data

The comprehensive characterization of somatic mutations by screening cancer genomes can help to understand cancer appearance and progression but also to identify accurately predictive biomarkers such as circulating tumor DNA (ctDNA). In 2015, the International Cancer Genome Consortium launched a large benchmarking operation with the objective of identifying and resolving issues of somatic mutation calling [1]. One conclusion of this study was that detecting somatic mutations in cancer genomes remains an unsolved problem. The problem is exacerbated when trying to identify low abundance mutations like in ctDNA due to reduced variant allelic fractions. Indeed, Next Generation Sequencing error level can reach this low proportion, and somatic variant calling from ctDNA is like finding a needle in a needlestack.

Here we present needlestack [2], a highly sensitive variant caller which learns from the data to precisely quantify the level of sequencing errors. needlestack is analyzing a large number of samples together to estimate the distribution of sequencing errors in order to accurately identify variants present in very low proportion. At each position and for each candidate alteration, i.e. single nucleotide substitutions and short insertions or deletions observed in the data, we model the sequencing error distribution using a negative binomial regression. Variants are detected as being outliers from this error model, and to avoid these outliers biasing the regression we adapted a recently published robust estimator for the negative binomial regression [3] based on a robust estimation of the overdispersion parameter. Needlestack can analyse a set of 50 whole exome sequenced samples in 20 hours when launched on 200 computing core.

We applied needlestack on 101 Small-Cell Lung Cancer (SCLC) cases and 306 non-cancer controls, to assess the presence of mutations in the cell-free DNA extracted from the plasma in the two most frequently mutated genes in these tumors: TP53 and RB1. We previously reported TP53 mutations in the cfDNA of 49% of SCLC cases and in 11.4% of controls [4], and using a genetic score using both genes increases the specificity to 96.2% for the same sensitivity. needlestack is implemented using Nextflow [5], allowing high scalability, portability and reproducibility by providing a Docker container image and versioned source code on GitHub [2].

References

[1] Alioto TS, Buchhalter I, Derdak S, et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun. 2015;6:10001.

[2] https://github.com/IARCbioinfo/needlestack

[3] Aeberhard WH, Cantoni E, Heritier S. Robust inference in the negative binomial regression model with an application to falls data. Biometrics. 2014;70(4):920-31.

[4] Fernandez-Cuesta L, Perdomo S, Avogbe PH, et al. Identification of Circulating Tumor DNA for the Early Detection of Small-cell Lung Cancer. EBioMedicine. 2016;10:117-123. doi:10.1016/j.ebiom.2016.06.032.

[5] P. Di Tommaso, et al. Nextflow enables reproducible computational workflows. Nature Biotechnology 35, 316–319 (2017) doi:10.1038/nbt.3820

CCB9. Fatemeh Dorri, Sean Jewell, Tyler Funnell, Alexandre Bouchard-Cote and Sohrab Shah
BC Cancer Research Center, Vancouver, Canada

MuClone: improved detection and classification of single nucleotide variants in multi-sample sequencing data

Accumulation of genetic alterations (e.g. single nucleotide variations, so called somatic mutations) in human cells is often causative of cancer initiation and progression. A complete catalogue of mutations across genome (numbering in the thousands) and their distribution across the cell populations that comprise a tumour is needed to decipher the clonal dynamics of human cancers. For this purpose, numerous somatic mutation-calling algorithms have been reported, but the problem remains challenging particularly for detecting low prevalence mutations which may be present in a small fraction of cells. These mutations have particular clinical importance as they may be low abundance at diagnosis, but increase in prevalence as a function of treatment and can act as indicators of therapeutic resistance.

Whole genome or exome sequencing from multiple samples in time or anatomic space of the same patient is an increasingly common experimental design for understanding the clonal composition of tumours. To better interpret multi-sample datasets, we developed a novel statistical framework, named MuClone, which simultaneously detects and classifies mutations across multiple tumour samples of a patient. MuClone's key advance is in incorporating prior knowledge about the cellular prevalences of tumour clones in a Bayesian statistical framework in order to improve the performance of detecting mutations- particularly those with low prevalence.

We used both synthetic and real data to evaluate the performance of MuClone. Analysis of synthetic data confirmed our hypothesis that injecting clonal population information into the inference of mutations from multiple samples improves the sensitivity of detecting somatic mutations without compromising specificity. This finding was robust over different values of tumour content and sequencing error rates and showed quantitative improvement over competing methods. We next performed analysis of multiple samples from patients with high-grade serous ovarian cancer, and multiple samples from patients with non–small-cell lung cancer. MuClone outperforms other methods in terms of Youden index on the validated mutations [1] from seven multi-site high-grade serous ovarian cancer datasets [1]. Performance of MuClone was statistically higher than MultiSNV based on Welch's t-test (p-value = 0.0006).

To further evaluate the performance of MuClone, we compared it against MultiSNV [2] on multiple whole exome sequencing samples of eight patients with non–small-cell lung cancer (from the TRACERx consortium) [3]. MultiSNV is one of the more efficient methods. MuClone outperformed MultiSNV notably on the list of reported mutations across multiple samples of different patients in TRACERx study. MuClone false negatives exhibited extremely low allelic ratios and there are only 19 false negative out of 7062 mutations (0.2%).

To interpret our results, we show that among the 59 (34) genes that were reported in TRACERx study for Adenocarcinoma (Squamous-carcinoma) samples, 49 (31) of them have recurrent mutations called by only MuClone. Furthermore, the Spearman's rank correlation coefficient between COSMIC mutation signatures on MuClone specific calls and TRACERx subclonal mutations resulted in high correlations between MuClone specific

mutation signatures and reported signatures (e.g. “Signature 13”, associated with activation of APOBEC cytidine deaminases and has been previously verified in [3, 4]).

In conclusion, we report that MuClone has increased accuracy for detecting mutations and that mutations found by MuClone, but no other methods show biological interpretability. As such, MuClone provides an advance in the field that will enhance the analysis of multi-sample tumour sequencing data.

References

[1] McPherson, A., et al. Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer, Nature Genetics (2016).
[2] Josephidou, M., et al. MultiSNV: a probabilistic approach for improving detection of somatic point mutations from multiple related tumour samples, Nucleic Acids Research (2015).
[3] Jamal-Hanjani, M., et al. Tracking the evolution of non-small-cell lung cancer, New England Journal of Medicine (2017).
[4] Campbell, J. D., et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas, Nature Genetics (2016).

CCB10. Faezeh Dorri, Fatemeh Dorri and Hector Corrada Bravo
University of Maryland, USA

Finding regions of differential methylation composition using bisulfite sequencing data

It is increasingly clear that cells within a population exhibit epigenetic heterogeneity, specifically exhibiting different cell-specific methylation profiles. Furthermore, when comparing cell populations across different phenotypic states, say in tumor vs. a normal tissue, a natural question to ask is what is the relationship between the heterogenous epigenetic cell-specific profiles found in the tumor cell population compared to the normal cell population. DNA methylation profiles can deviate from normal cell populations by either the inclusion of new cell-specific methylation patterns or changes in the abundance of cell-specific methylation patterns also found in the normal cell population.

In this work, we propose a statistical model that leverages our previous work on cell-specific methylation pattern reconstruction to understand the role of epigenetic heterogeneity at the cell population level. In our model, we first use the BS-seq read counts to reconstruct the underlying cell-specific methylation patterns across genomes. Then, we exploit the composition of inferred methylation patterns from the first step in a same cell-type population to capture the variation in composition of patterns. We introduce the notion of regions of differential methylation composition (RDMCs) by defining a statistic that considers the differences in the distribution of cell-specific methylation patterns across tumor and normal cell populations. The significance of each RDMCs is assessed via a permutation procedure and experimental results confirm the improvement in the sensitivity and specificity of our model compared with other methods designed for differential composition analysis in other genomic applications.

CCB11. Aurélie Gabriel, Nicolas Alcala, James McKay, Lynnette Fernandez-Cuesta and Matthieu Foll
Genetic Cancer Susceptibility Group, Section of Genetics, International Agency for Research on Cancer (IARC-WHO), Lyon, France

Multi-omics pan-cancer classification using machine learning

Thanks to initiatives like The Cancer Genome Atlas (TCGA) or the International Cancer Genome Consortium (ICGC), scientists have access to different 'omics data generated from multiple cancer types. Several studies take advantage of these data to perform pan-cancer classification but often focus on one data category rather than combining multiple types of data (WXS, RNAseq, CNVs, methylation …). A multi-omics approach could increase the performance of such classifiers, allow the identification of the most informative features for the classification of different cancer types and provide guidelines for biomarkers design. In this study, we first present a pan-cancer multi-omics classification analysis of the TCGA data using machine learning algorithms (support vector machines, random forest). The training and parameter tuning were done on a training and validation set using a 10 fold cross validation and an independent test set was saved to test and evaluate performance. We compare the performance of classifiers based on different molecular features: mutated driver genes, mutational burden, distribution of nucleotides changes, copy number variation, fusion transcripts, expression and methylation. The features were selected based on prior biological knowledge of each cancer to reduce dimensionality. As previously shown expression and methylation data lead to a higher performance. However, this information is difficult to translate into the clinical setting. In this context, our data suggest that features derived from mutation, copy number and fusion transcript data can achieve a good performance in many cases, which could be more easily translated in a clinical context. As proof-of-principle example, we use the same approach to distinguish one particular cancer type from any others and apply feature selection methods to identify the smallest number of features needed for this task. We apply this to develop a biomarker for the early detection of small cell lung cancer and show that a handful of features including the two most frequently mutated genes, TP53 and RB1, can lead to a classifier with a good sensitivity and specificity.

CCB12. M.C. Baldauf[1,*], J.S. Gerke[1,*], A. Kirschner[2], F. Blaeschke[3], M. Effenberger[4], K. Schober[4], R. Alba Rubio[1], T. Kanaseki[5], M.M. Kiran[6], M. Dallmayer[1], J. Musa[1], N. Akpolat[6], A.N. Akatli[6], F.C. Rosman[7], Ö. Özen[8], S. Sugita[5], T. Hasegawa[5], H. Sugimura[9], D. Baumhoer[10], M.M.L. Knott[1], G. Sannino[1], A. Marchetto[1], J. Li[1], D.H. Busch[4], T. Feuchtinger[3], S. Ohmura[1], M.F. Orth[1], U. Thiel[2], T. Kirchner[11,12,13], T.G.P. Grünewald[1,11,12,13,§]
1. Max-Eder Research Group for Pediatric Sarcoma Biology, Institute of Pathology of the LMU Munich, Munich, Germany; 2. Children's Cancer Research Center, Technische Universität München, Munich, Germany; 3. Dr. von Hauner'sches Children's Hospital, LMU Munich, Munich, Germany; 4. Institute for Medical Microbiology, Immunology and Hygiene, TU Munich, Munich, Germany; 5. Sapporo Medical University, Sapporo, Japan; 6. Department of Pathology, Turgut Ozal Medical Center, Inonu University, Malatya, Turkey; 7. Hospital Municipal Jesus, Rio de Janeiro, Brazil; 8. Başkent University Hospital, Turkey; 9. Hamamatsu School of Medicine, Hamamatsu, Japan; 10. Bone Tumor Reference Center, Institute of Pathology of the University Hospital of Basel, Switzerland; 11. Institute of Pathology of the LMU Munich, Munich, Germany; 12. German Cancer Consortium (DKTK), partner site Munich, Heidelberg, Germany; 13. German Cancer Research Center (DKFZ), Heidelberg, Germany

Systematic identification of cancer-specific immunogenic peptides with RAVEN

Immunotherapy can revolutionize anti-cancer therapy if specific targets are available. Recurrent somatic mutations in the exome can create highly specific neo-antigens. However, especially pediatric cancers are oligo-mutated and hardly exhibit recurrent neo-antigens. Yet, immunogenic peptides encoded by cancer-specific genes (CSGs), which are virtually not expressed in normal tissues, can enable a targeted immunotherapy of such cancers.

Here, we describe an algorithm and provide a user-friendly software named RAVEN (Rich Analysis of Variably Expressed genes in Numerous tissues), which automatizes the systematic and fast identification of CSG-encoded peptides highly affine to Major Histocompatibility Complexes (MHC) based on publicly available gene expression data.

We applied RAVEN to a data set assembled from more than 2,600 simultaneously normalized gene expression microarrays comprising 50 tumor entities, with a focus on sarcomas and pediatric cancers, and 71 normal tissue types. RAVEN performed a transcriptome-wide scan in each cancer entity for gender-specific CSGs. As a proof-of-concept we identified several established CSGs, but also many novel candidates that are applicable for targeting multiple cancer types. The specific expression of the most promising CSGs was validated by qRT-PCR in cancer cell lines and by immunohistochemistry in a comprehensive tissue-microarray.

By implementing an artificial neural network algorithm and by crosschecking with the UniProt protein-database, RAVEN subsequently identified unique CSG-encoded peptides with high affinity to MHCs and without sequence similarity to abundantly expressed proteins. The predicted MHC-affinity of these peptides was validated in T2-cell peptide-binding assays in which many showed similar kinetics to an extremely immunogenic influenza control peptide.

Collectively, we provide a comprehensive, exquisitely curated and validated catalogue of cancer-specific and highly MHC-affine peptides across 50 cancer entities. Additionally, we implemented an intuitive software to examine any gene expression data set with our developed algorithm and methods (https://github.com/JSGerke/RAVENsoftware). We anticipate that our peptide catalogue and software will constitute a rich resource to accelerate immunotherapy development.

CCB13. Shila Ghazanfar, Dario Strbenac, John T. Ormerod, Jean Yee Hwa Yang and Ellis Patrick

Judith and David Coffey Lifelab, School of Mathematics and Statistics, University of Sydney, Australia

DCSR: Differential correlation across survival ranking

See abstract at p. 10.

CCB14. Ariel Afek[1], Raluca Gordan[1,2]

1. Center for Genomic and Computational Biology, Department of Biostatistics and Bioinformatics, 2. Departments of Computer Science, Department of Molecular Genetics and Microbiology, Duke University, Durham, USA

Widespread increase in transcription factor-DNA binding due to mismatch damage

DNA mismatches occur when two non-complementary bases are aligned on opposite strands of a DNA duplex, forming a 'mispair'. They are generated during DNA replication, genetic recombination, and by frequent spontaneous DNA deamination. Mismatches alter the DNA structure and the functional groups available in the DNA major/minor grooves, which can affect interactions with regulatory transcription factors (TFs). Currently, very little is known about the effects of mismatches on TF binding. We present Saturation Mismatch Binding Assay (SaMBA), the first assay to characterize the effects of mismatches on TF-DNA binding in high-throughput. For a set of genomic sequences of interest, SaMBA generates DNA duplexes containing all possible single-base mismatches, and quantitatively assesses the effects of the mismatches on the binding specificity of TFs. We applied SaMBA to measure binding of 18 human and yeast TFs (covering 10 distinct structural families) to thousands of mismatched DNA sequences. Interestingly, for all tested factors we found that DNA mismatches within TF binding sites can significantly increase TF binding levels compared to the wild-type sequences. The magnitude of this effect is large, with some mismatches leading to a >5-fold increase in TF binding signal measured in our assay. In addition, we found that mismatches can increase the TF binding level even for the highest affinity binding sites currently known. We show that the effect of single base mismatches is different from that of single base-pair mutations, and that both DNA base readout and DNA shape readout likely contribute to the increased binding affinity caused by single base mismatches. Furthermore, for several TFs we have identified genomic regions with no putative binding sites that become strongly bound after certain mismatches are introduced. For example, we identified a nonspecific DNA site for which certain mismatches (which can result from 5-Methylcytosine deamination) lead to a ~30-fold increase in binding of c-Myc, resulting in a binding affinity higher than native E-box binding sites. The non-specific site occurs ~11,000 times in human promotes, and spontaneous deamination is estimated to occur 100-500 times per cell per day in humans, making such nonspecific sites candidates to become new TF targets in human promoters. In conclusion, increased binding of eukaryotic TFs due to DNA mismatches is a widespread phenomenon, as we observed it for all tested proteins. The high affinity of TFs for sites with certain DNA mismatches has the potential to influence gene expression and DNA repair processes, especially in mismatch repair-deficient cancer cells.

CCB15. Yuanhua Huang, Davis J. McCarthy, Raghd Rostom and Oliver Stegle

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

Cardelino: clonal assignment of single cells with expressed mutations

See abstract p.16.

CCB16. Kim Hirshfield, Mendel Goldfinger, Simon Bird, Mohammad Hadigol, Sepand Ansari, Lorna Rodriguez-Rodriguez, Shridar Ganesan and Hossein Khiabanian

Rutgers Cancer Institute of New Jersey, USA

Inference of germline mutational status and of loss of heterozygosity in high-depth tumor-only sequencing data

Inherited germline defects are implicated in up to 10% of human tumors, with particularly well-known roles in breast and ovarian cancers that harbor BRCA1/2-mutated genes. There is also increasing evidence for the role of germline alterations in other malignancies such as colon and pancreatic cancers. Mutations in familial cancer genes can be detected by high throughput sequencing (HTS), when applied to formalin-fixed paraffin-embedded (FFPE) tumor specimens. However, due to often lack of patient-matched control normal DNA and/or low tumor purity, there is limited ability to determine the genomic status of these alterations (germline versus somatic) and to assess the presence of loss of heterozygosity (LOH). These analyses, especially when applied to genes such as BRCA1/2, can have significant clinical implications for patient care, which often cannot be answered by routine germline testing. Here, we present LOHGIC (LOH-Germline Inference Calculator), developed on a model-selection scheme using Akaike Information Criterion weighting. LOHGIC infers the most consistent model describing the germline-versus-somatic mutational status, and predicts LOH for mutations identified via clinical grade, high-depth, hybrid-capture tumor-only sequencing. It also incorporates statistical uncertainties inherent to HTS as well as biases in tumor purity estimates. We used LOHGIC to assess BRCA1/2 mutations in 1,636 specimens sequenced at Rutgers Cancer Institute of New Jersey. Evaluation of LOHGIC with available germline sequencing from BRCA1/2 testing demonstrated 93% accuracy, 100% precision, and 96% recall. This analysis also highlighted a differential tumor spectrum associated with BRCA1/2 mutations. LOHGIC, available at software.khiabanian-lab.org, can be applied to any gene with candidate, inherited mutations. This approach demonstrates the clinical utility of targeted sequencing in both identifying patients with potential germline alterations in tumor suppressor genes as well as estimating LOH occurrence in cancer cells, which may confer therapeutic relevance.

CCB17. Sarah Killcoyne, Eleanor Gregson, David Wedge, Matthew Eldridge, Moritz Gerstung and Rebecca Fitzgerald

European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

Early detection of cancer progression using whole-genome copy number in pre-malignant esophageal tissues

At the genomic level esophageal adenocarcinomas (EAC) are predominately driven by copy-number alterations rather than single nucleotide mutations. Both small and large-scale alterations can be observed, though their specific relevance to the progression of the disease is unclear. A pre-malignant tissue, known as Barrett's Esophagus, has been shown to develop early somatic copy number changes that may persist and increase in complexity in EAC providing window into the development of these tumors. However, the risk for progressing from Barrett's to EAC is only around 0.3% per year. This has significant implications for both our understanding of the tumor development as well as clinical management of such patients.

In a nested case-control cohort of 89 patients diagnosed with Barrett's esophagus split into two groups based on their progression to early cancer, we investigated the copy number status of patients over time. The resulting 793 samples across all patients was collected from clinical biopsies from up to 15 years of patient visits. Shallow whole-genome sequencing (0.4x) was performed on all samples, and the copy number status derived using publicly available tools. The copy number information was then used to develop a computational model to predict progression on a per-sample and per-endoscopy basis (e.g. single clinical visit).

Using nested leave-one-out prediction, the model robustly classifies both progressive and non-progressive samples with an AUC of 0.9. At the pathological endpoint the model predicted 92-96% of the early cancer samples correctly as progressive, but most important were the predictions at pathologies prior to the endpoint where current clinical diagnosis is not performed. The model predicted 77% of non-dysplastic Barrett's samples as progressive in patients that later developed cancer. It is also predicted progression in 75-90% of endoscopies for progressive patients up to 10 years prior to diagnosis. These predictions were validated both using in-silico methods, and a separate cohort of Barrett's tissues from patients with confirmed EAC. Using this model we are able to provide patient treatment recommendations through an application using copy number data that is practical and simple to generate in a clinical context.

CCB18. Yoo-Ah Kim[1*], Damian Wojtowicz[1], Annie Zhao[2], and Teresa M. Przytycka[1]

1. National Center of Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD, USA 2. Montgomery Blair High School, Silver Spring, MD, USA

Implications of non-B DNA structures in cancer mutagenesis

Recent advances in systematic analysis of whole genome cancer mutations revealed that cancer patients undergo combinations of mutational processes, accumulating different patterns of somatic mutations as a byproduct of such processes. Dozens of such patterns (named mutational signatures) were uncovered through decomposition of mutation catalogues in cancer genomes and linked with mutagenic processes such as defective DNA maintenance, UV light exposure, APOBEC activities, tobacco consumption etc. On the other hand, several evidences about the role of non-B DNA structures in cancer mutagenesis have been reported. For example, genomic regions with short inverted repeats tend to form a secondary structure called a hairpin and may confer increased localized mutability. In this work, we interrogated the whole genome mutational profiles of more than 3000 cancer patients from 22 different cancer types and assessed the enrichment of non-B DNA structures to better understand their implications in cancer mutagenesis.

CCB19. Nan Li[1,2], Roman Schefzik[1], Bernd Fischer[1], Ângela Teresa Gonçalves Filimon[1,3]

1. German Cancer Research Center, Heidelberg, Germany; 2. University of Heidelberg, Heidelberg, Germany; 3. Wellcome Trust Sanger Institute, Cambridge, UK

Stepwise Elastic Net Regression with group penalties in anti-cancer drug response prediction with multiple genomic data

The identification of biomarkers that can predict patients' response to anti-cancer drugs is a topic of great interest in personalised oncology. In recent years, several projects have tested sets of drugs on panels of human cancer cell lines for which multiple genomic variables have been recorded. Previous work using Elastic Net regression (ENR) has had only limited success due to widely differing dimensionalities between data types, as well as the intricate correlation among them.

To address these problems, we have developed a method called stepwise ENR, which accounts for different data types in a successive manner. Specifically, for each drug, we: i) fit drug response with one single data type, then fit the residuals to another data type, repeat until all data have been fit; and ii) randomise the order of data types.
This approach allows us to estimate the performance of each data type in the prediction, with and without receiving information from all other data types. To achieve this, we introduce the definition of separated correlation, which is the total information of one single data type received from all other data types.

Using our model we have observed that in the CCLE data set, for 23 out of 24 drugs, the separated correlation is not ignorable, demonstrating the deficiency of classic ENR. The important features for prediction differ from drug to drug. For example, for the EGFR-targeted drug Erlotinib, gene expression and cancer type are important features, which gives us a hint in understanding the mechanism of Erlotinib.

CCB20. Ariel Afek[1,2*], Zachery Mielko[1,3*], Debbie Burdinski[1,2], Sheera Adar[4], Raluca Gordan[1,2,5]
1. Center for Genomic and Computational Biology, Duke University, Durham, USA; 2. Department of Biostatistics and Bioinformatics, Duke University, Durham, USA; 3. University Program in Genetics and Genomics, Duke University, Durham, USA; 4. Department of Microbiology and Molecular Genetics, The Hebrew University, Jerusalem, Israel; 5. Department of Computer Science, Department of Molecular Genetics and Microbiology, Duke University, Durham, USA; *Co-first authors

The impact of UV damage on transcription factor binding specificity

Ultraviolet (UV) radiation is a major carcinogen for most skin cancers. UV is absorbed in DNA and leads to formation of specific di-pyrimidine (TT/TC/CC/CT) photoproducts. Recent studies suggest transcription factors (TFs) can hinder the repair of UV damage in TF binding sites, which subsequently leads to mutations (Sabarinathan, et al., 2016). However, very little information is available on how/whether transcription factors bind to UV-damaged DNA sites. To assess the impact of UV on TFDNA binding specificity, we developed a new high-throughput assay that comprehensively measures *in vitro* transcription factor binding to UV-damaged DNA.

Our assay is based on the widely-used protein binding microarray (PBM) technology, but instead of the 'universal' designs of (Berger and Bulyk, 2009) we utilize the UV-damageable site composition to specifically design a DNA library that avoids damage at off-target sites. The library is synthesized de novo onto Agilent glass slides and double-stranded by primer extension. Afterwards, we induce UV damage to the DNA oligos on the slide by exposing it to UV light. The amount of UV damage can be controlled through exposure time and amount of radiation (we use a Stratalinker instrument and a custom setup for the slides).

We chose two human TFs, c-Myc and Ets1, as our case studies to measure and compare TF-DNA binding specificities before and after UV exposure. UV damage is well characterized, creating cyclobutanepyrimidine dimers (CPD) and 6-4 photoproducts (6-4PP), which distort the shape of DNA. We confirmed UV damage formation in each sequence using CPD and 6-4PP antibodies, both of which resulted in highly reproducible data (squared Pearson correlation coefficient between replicates: R2 = 0.98 for CPD and R2 = 0.98 for 6-4PP). The composition of damage types was consistent with *in vivo* data (Hu et al 2017).

Our results illustrate the changes in DNA-binding specificity caused by UV damage. For example, we find that after UV exposure, Ets1 preferentially binds to (A/G)TCCGG instead of TTCCGG sites, avoiding the highly damageable TT dimers. This suggests that UV conditions alter the genomic binding profiles of TFs, which in turn can affect the TFs' competition with repair enzymes, and subsequently the formation of mutations.

CCB21. Hanna Najgebauer[1,2], Mi Yang[3], Hayley Francies[4], Euan A Stronach[1,5], Julio SaezRodriguez[1,2,3], Mathew J Garnett[1,4], Francesco Iorio[1,2,4]

1. Open Targets, Wellcome Genome Campus, Cambridge, UK; 2. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK; 3. Faculty of Medicine, Joint Research Centre for Computational Biomedicine, RWTH Aachen University, Germany; 4. Wellcome Trust Sanger Institute, Wellcome Genome Campus, Cambridge, UK; 5. Target Sciences, GlaxoSmithKline, Stevenage, UK

CELLector: Genomics Guided Selection of Cancer *in vitro* Models

See abstract p. 11.

CCB22. Francesca Petralia[1,2,*], Li Wang[1,2,3,*], Jie Peng[4], Arthur Yan[1,2], Jun Zhu[1,2,3] and Pei Wang[1]

1. Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, USA; 2. Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, USA; 3. Sema4, a Mount Sinai venture, Stamford, CT, USA; 4. Department of Statistics. University of California, Davis, CA, USA

A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity

Tumor tissue samples often contain an unknown fraction of stromal cells. This problem well known as tumor purity heterogeneity (TPH) was recently recognized as a severe issue in omics studies. Specifically, if TPH is ignored when inferring co-expression networks, edges are likely to be estimated among genes differentially expressed between non-tumor and tumor cells rather than among gene pairs interacting with each other in tumor cells. To address this issue, we propose TSNet a new method which constructs tumor-cell specific gene co-expression networks based on gene expression profiles of tumor tissues. TSNet treats the observed expression profile as a mixture of expressions from different cell types and explicitly models tumor purity percentage in each tumor sample. In practice, tumor-purity can be estimated using other data types such as DNA copy number profiles. However, as shown by [1], DNA and RNA data result in similar tumor-purity levels for several cancer types and, therefore, expression data can be directly utilized to estimate both tumor purity

and co-expression networks. In particular, TSNet is designed to borrow information from existing prior information when inferring tumor-purity levels from gene expression data. Once that tumor-purity is estimated, co-expression networks for both tumor and non-tumor components are inferred. Using extensive synthetic data experiments, we demonstrate the advantage of TSNet over the standard graphical lasso (mixNet) [2] which does not take into account tumor-purity heterogeneity. For this synthetic data scenario, network topologies for both tumor and non-tumor components were randomly sampled from the power law distribution. The two algorithms were compared counting the number of false positive and true positive edges in estimating tumor-specific and normal-specific networks on synthetic data involving different number of observations (n = 200, n = 400). TSNet results in higher true positive rates and lower false positive rates for all simulation scenarios. We then apply TSNet to estimate tumor specific gene co-expression networks for TCGA ovarian cancer RNAseq data. We show that our tumor-purity estimates are comparable to those of other two well-known algorithms for tumor-purity estimation [3,4]. After applying TSNet to the RNAseq data of 281 TCGA ovarian tumor samples, we construct gene co-expression networks for the tumor and non-tumor components, respectively. We identified the top genes highly connected in TSNet-networks and poorly connected in the network estimated using the standard graphical lasso. In particular, the leading gene of this type, HIC1, is a tumor-repressor gene which plays an important role in both TSNet-Tumor and TSNet-Normal networks. Further investigation of HIC1 neighbors in TSNet-Tumor and TSNet-Normal revealed that this gene is involved in different biological processes in the tumor and non-tumor components. In fact, the neighborhood of HIC1 in TSNet-Tumor is enriched of DNA damage pathways such as "Reactome p53 independent DNA damage response"; while the neighborhood in TSNet-Normal is enriched of ribosomal and immune system related pathways. These findings are well supported by the recent literature on HIC1. To further evaluate the performance of TSNet, we carried out an enrichment analysis. Specifically, we observe that highly connected genes and network neighbors in TSNet-Tumor were enriched of DNA damage pathways such as "Reactome DNA Repair" and other cancer related signaling pathways such as "Kegg MAPK Signaling Pathway" and "Kegg ERBB Signaling Pathway"; while those in TSNet-Normal were enriched of stromal genes and extracellular matrix pathways. These results again illustrate the capability of TSNet to accurately characterize biological activities specific to tumor and non-tumor tissues, which could be very useful for studying gene interaction mechanisms underlying diseases.

CCB23. Nimrod Rappoport and Ron Shamir
School of Computer Science, Tel Aviv University, Israel

MOCCASIN: Cancer subtyping by multi-omics integration

Recent technological advances have facilitated the production of large high throughput biological data types, collectively termed "omics". These include genomics, transcriptomics, proteomics and many more. Analysis of omics datasets has proven invaluable for basic biological research and for medicine. Research in computational biology initially focused on analyzing each omic type separately. While inquiry of each data type separately provides insights on its own, integrative analysis of multiple data types may reveal more holistic, systems-level insights.

The large, diverse omics data available today can be used to characterize human disease better, and to help physicians treat patients in a more personalized way. In oncology, analysis of large datasets has led to the discovery of novel cancer subtypes. The classification of tumors into these subtypes is now used in treatment decisions. However, these subtypes are usually defined through the use of a single omic (e.g. gene expression).

Using multiple omics for cancer subtyping will allow us to better understand cancer biology, and to suggest more effective and precise therapy.

In this work we present MOCCASIN (Multi Omic Clustering CAptured by SImilarity of Neighborhoods), a novel algorithm for cancer subtyping through integration of several omic datasets. By using similarities between patients, the algorithm can handle diverse omics without having to model each omic separately, and can support omics with hundreds of thousands of measurements per patient.

MOCCASIN is fast and simple, and has the added advantage of handling missing data, i.e. it can include in the analysis patients for which not all types of data are available. Preliminary results on several cancers show that MOCCASIN partitions the tumors into groups that are distinctive in terms of prognosis patterns and other clinical parameters, even in missing data situations. Those results are comparable with state-of-the-art algorithms on full datasets, and show an improvement over extant algorithms that handle missing data.

CCB24. Sabrina Rashid, Sohrab Shah, Ziv Bar-Joseph and Ravi Pandya
Computational Biology Department, Carnegie Mellon University, Pittsburgh, USA

Dhaka: Variational autoencoder for unmasking tumor heterogeneity from single cell genomic data

See abstract p. 14.

CCB25. Michele Caselle, Francesca Orso, Daniela Taverna, Laura Cantini, Loredana Martignetti, Matteo Osella, Antonio Rosanova and Elisa Reale
University of Turin, Italy

Searching for epi-miRNA using transfection experiments

MicroRNAs which regulate epigenetic factors are usually denoted as epi-miRNA. There is growing evidence that these epi-miRNAs play an important role in several differentiation processes and more generally that in higher eukaryotes there is a strong interplay between post-transcriptional regulation, mediated by microRNAs, and epigenetic regulation. In this work we show how to use existing miRNA transfection experiments to obtain lists of candidate epi-miRNAs and characterize their main biological properties. The procedure that we propose can be used both to prioritize candidate epi-miRNAs for further experiments and, given a microRNA of interest, to identify its candidate epigenetic interactors. We study, as an example, the case of mir-214, identify as putative interactor EZH2 which is a component of the Polycomb repressive complex 2 and discuss its role in a cell line of melanoma.

CCB26. Camir Ricketts, Daniel Seidman, Victoria Popic, Fereydoun Hormozdiari, Serafim Batzoglou, Iman Hajirasouliha
Weill Cornell Medicine, New York, USA

Meltos: Multi-sample tumor phylogeny reconstruction for structural variants

See abstract p. 8.

CCB27. Mohammed El-Kebir[1,3], Gryte Satas[2], Benjamin J. Raphael[1,*]
1. Department of Computer Science, Princeton University, Princeton, NJ, USA; 2. Department of Computer Science, Brown University, Providence, RI, USA; 3. Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

Inferring parsimonious migration histories for metastatic cancers

Recent studies have attempted to infer the pattern of cellular migrations between a primary tumor and distant metastases using phylogenetic trees constructed from somatic mutations. Several of these studies have challenged the conventional view of monoclonal seeding of a metastasis, and reported complex patterns of migration between primary tumors and metastases, including polyclonal seeding and reseeding.

Most current phylogenetic analyses that attempt to infer migration patterns from somatic mutations rely on two key assumptions, sample homogeneity and mutation-migration concordance, that do not typically hold for cancer sequencing datasets. Consequently, standard phylogenetic techniques may result in incorrect or misleading patterns of metastasis.

We introduce a computational model to rigorously evaluate migration patterns. We use this model as a basis for Metastatic And Clonal History INtegrative Analysis (MACHINA), an algorithm that jointly infers parsimonious clone trees and migration histories of metastatic cancers from DNA sequencing data. We show that MACHINA accurately recovers clone trees and migration histories on simulated data. We apply MACHINA to sequencing data from metastatic ovarian, breast, prostate and skin cancer samples. In several cases, we find simpler migration histories than previously reported, altering conclusions regarding metastasis-to-metastasis spread, the anatomical site of the primary tumor, or the occurrence of polyclonal seeding and reseeding in individual patients. MACHINA enables researchers to rigorously assess the validity of different migration patterns in individual patients with metastatic disease and evaluate the prevalence of different migration patterns across large cohorts of patients and tumor types.

CCB28. Roman Schulte-Sasse and Annalisa Marsico
Max Planck Institute for Molecular Genetics, Dresden, Germany

Predicting cancer genes with graph convolutional networks

Despite the vast increase of data from next-generation sequencing and the ever-growing knowledge about the functionality of genes, the prediction of genes involved in disease remains challenging. This is partly due to the complexity of diseases like cancer but also due to the lack of efficient algorithms that make use of knowledge from different sources and domains for their predictions.

Here, we present a powerful classifier, the graph convolutional network (GCN) in the context of disease gene prediction and show how it can combine protein-protein interaction (PPI) networks and gene expression data to predict cancer-related genes. The algorithm can use the two different kinds of knowledge to infer new disease genes and is embedded in a deep learning setting. Therefore it can handle multi-class prediction and vectors of features for each gene while competing solutions can either not make use of the PPI network or are bound to only use scalar input features.

We show that our model does not suffer from the problem that it only predicts prominent, high-degree genes but also less connected genes that are biologically relevant.

We compare our method with Hotnet2, a popular tool for extracting highly mutated subnetworks and NetRank, a PageRank variant. However, both methods can only use scalar features and are based on random walks over the PPI network. After training our deep network, we extract the features that contribute most to the classification result for each gene to find underlying mechanisms in network topology and the features.

CCB29. Jesse Eaton, Jingyi Wang and Russell Schwartz
Biological Sciences, Carnegie Mellon University, Pittsburgh, USA

Joint deconvolution and phylogeny inference of tumor structural variation data

Introduction: Cancer genomic research has made it possible to observe extensive genetic variation patient-to-patient (intertumor heterogeneity) and cell-to-cell within single patients (intratumor heterogeneity) and revealed a complex landscape of somatic variations. Reconstructing how individual tumors progress from these large and highly stochastic data sources, however, has proven daunting. A key insight into interpreting tumor genomic data was tumor phylogenetics [1], i.e., the use of phylogenetic inference to reconstruct tumor progression, which arose to take advantage of the fact that tumor progression is driven by evolution at the cellular level and that it might thus in principle be understood by algorithms for reconstructing evolutionary trees. Nonetheless, this general intuition can only take one so far because tumor evolution is quite different in its details from the species evolution assumed by most standard phylogeny methods, leading to a large literature of novel methods for tumor phylogeny inference for varying scales, study designs, and data sources. Most tumor genomic data today is in the form of bulk sequence mixing together distinct cell populations and thus tumor phylogeny methods are commonly joined to methods for genomic deconvolution, as in, for example, THeTA [2]. Such methods almost exclusively work using either single nucleotide variants (SNVs), copy number aberrations (CNAs), or, more recently combinations of the two [3]. An important limitation of prevailing methods for tumor deconvolution and phylogenetics is an inability to work with general structural variations (SVs), which are crucial to tumor progression and phenotypic adaptation, for example via induced CNAs or the creation of novel gene fusions. We introduce a method for deconvolution of broader classes of SVs, including segmental duplications, deletions, inversions, and translocations inferred from mate paired whole genome sequence (WGS) data.

Methods: At a high level, our method relies on a constraint satisfaction framework to simultaneously deconvolve clonal populations from bulk variation data and fit them to a tumor phylogeny. The method takes as input a set of SV data for one or more genomic samples from a tumor, in the form of a matrix F of average copy numbers across genomic segments, a matrix Q mapping SV breakpoints to genomic segments, and a matrix G identifying mated pairs of breakpoints. For the present implementation, we assume these inputs are inferred with Weaver [4], a program for calling SVs at single-nucleotide resolution from WGS data. It uses these inputs to infer a clonal structure U,C where C represents copy numbers inferred for clonal cell populations and U describes how these cell populations are proportioned among samples, optimizing for an objective function 〖min〗_(U,C) (|F-UC|+λ_1 R+λ_2 S) (Eq. 1) where R represents an L1 cost of a phylogenetic tree describing the evolution of CNAs among inferred clones and S is a cost capturing consistency of copy numbers of breakpoints

and the segments containing them. λ1 and λ2 are regularization terms, estimated empirically from the input data. We estimate λ1 as (l+r)(m)/(Nl) and λ2 as (l+r)/l for data with l breakpoints, r segments, m samples, and N phylogeny nodes to be inferred.

The objective function is optimized relative to a set of linear constraints enforcing consistency between clonal structure, the inferred phylogenetic tree R, and the input data. The complete model is specified as an integer linear program (ILP). This is made possible by a limited perfect phylogeny constraint, which assumes that nucleotide-resolution SVs are not subject to recurrent mutation, but which allows for subsequent loss or copy number variation without restriction. This assumption allows us to take advantage of a relaxed version of the ancestry condition of [5] in order to limit the solution search space. (We omit the full integer linear program (ILP) as it will not fit within the abstract page limit.)
We solve for the objective by a heuristic coordinate descent algorithm. In this algorithm, we iterate between optimal solutions of U given C and C given U, starting from a randomly initialized U. The iteration is repeated until convergence or a present maximum number of iterations. We repeat this for a user-selected number of random restarts (10 for simulated data, 2 for real data below).

Results: We first validated the methods with simulated data, generating assorted scenarios of mutations, samples, and numbers of clones and placing mutations, with SVs accumulated randomly by a Poisson model. Mutations were selected uniformly across three SV types (duplication, deletion, inversion) with exponentially distributed sizes. Average results showed a substantial improvement in accuracy of inference of clonal copy number vectors for the model of equation 1 (mean L1 distance 450) relative to a baseline model solving only for the first (deconvolution) term of the objective (mean L1 distance 760). These results show that the phylogenetic SV constraints lead to substantially improved deconvolution accuracy relative to deconvolution naïve of SVs and phylogenetic cost.

We further verified the effectiveness of the methods on a selection of TCGA breast cancer BRCA [6] samples for which WGS data was available. Of 59 samples, 31 ran to completion within two days of analysis on a 128 Gb computer, with the remainder failing due to excessive runtime or memory overflow. This suggests that the method is sufficiently efficient to handle a majority of real WGS tumor data sets, although it is challenged by those with the highest SV counts. While we cannot know the ground truth of these samples, the resulted in diverse topologies and tree complexities, generally supportive of highly stochastic, branched, and patient-specific trajectories of tumor evolution.

Discussion: We have developed a model and coordinate-descent ILP algorithm for joint deconvolution and phylogeny inference of tumor genomic data meeting a need for SV-aware tumor phylogenetics. We demonstrate that the method significantly improves deconvolution accuracy relative to baseline CNA deconvolution not accounting for SV phylogenetics. Application to TCGA BRCA samples show it to be able to handle a majority of real tumor datasets, although needed compute resources are still prohibitive for samples with the highest SV rates. The work provides a first step towards SV-aware deconvolution and phylogeny inference, although important work remains to improve both the models and algorithms and apply them more extensively to analysis of mechanisms of SV-driven evolution in cancers.

Acknowledgments: We thank Ashok Rajaraman and Jian Ma for helpful discussions and assistance with Weaver. Portions of this work were funded by U.S. National Institutes of Health award R21CA216452 and Pennsylvania

Department of Health Grant GBMF4554 #4100070287. The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). The results published here are in whole or part based upon data generated by The Cancer Genome Atlas managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

References:

[1] R. Desper et al. J. Comput. Biol. 6.1 (1999), pp. 37–51.
[2] L. Oesper, G. Satas, & B. J. Raphael. (2014). Bioinformatics, 30(24), 3532-3540.
[3] A. G. Deshwar et al. Genome Biol. 16 (2015), p. 35.
[4] Y. Li et al. Cell 3 SEP (2016).
[5] M. El-Kebir et al. Bioinformatics, 31(12), pp.i62-i70.
[6] Cancer Genome Atlas Network. Nature 490.7418 (2012), pp. 61–70.

CCB30. Linda K. Sundermann[1,2,*], Daniel Doerr[2], Amit G. Deshwar[3,4], Jeff Wintersinger[5], Jens Stoye[2], Quaid Morris[3,5,6,*] and Gunnar Rätsch[7,8,*]

1. International Research Training Group GRK 1906/1; 2. Genome Informatics, Faculty of Technology and Center for Biotechnology, Bielefeld University, Germany; 3. Edward S. Rogers Sr. Department of Electrical and Computer Engineering, University of Toronto, Canada; 4. Deep Genomics Inc., Canada; 5. Department of Computer Science, University of Toronto, Canada; 6. The Donnelly Center for Cellular and Biomolecular Research, University of Toronto, Canada; 7. Biomedical Informatics, Department of Computer Science, ETH Zürich, Switzerland; 8. Computational Biology Center, Memorial Sloan Kettering Cancer Center, New York, USA; *Corresponding authors: lsunderm@cebitec.uni-bielefeld.de, quaid.morris@utoronto.ca, gunnar.ratsch@ratschlab.org

Onctopus: Lineage-based subclonal reconstruction

See abstract p. 6.

CCB31. Mamoru Kato[1], Daniel A. Vasco[2,†], Ryuichi Sugino[3], Daichi Narushima[1] and Alexander Krasnitz

1. Department of Bioinformatics, National Cancer Center Research Institute, Tokyo, Japan; 2. Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, USA; 3. School of Advanced Sciences, The Graduate University for Advanced Studies, Hayama, Japan; 4. Cold Spring Harbor Laboratory, Simons Centre for Quantitative Biology, New York, USA

Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer

Single-cell sequencing is a promising technology that can address cancer cell evolution by identifying genetic alterations in individual cells. In a recent study, genome-wide DNA copy numbers of single cells were accurately

quantified by single-cell sequencing in breast cancers. Phylogenetic-tree analysis revealed genetically distinct populations, each consisting of homogeneous cells. Bioinformatics methods based on population genetics should be further developed to quantitatively analyse the single-cell sequencing data. We developed a bioinformatics framework that was combined with molecular-evolution theories to analyse copy-number losses. This analysis revealed that most deletions in the breast cancers at the single-cell level were generated by simple stochastic processes. A non-standard type of coalescent theory, the multiple-merger coalescent model, aided by approximate Bayesian computation fit well with the data, allowing us to estimate the population-genetic parameters in addition to false-positive and false-negative rates. The estimated parameters suggest that the cancer cells underwent sweepstake evolution, where only one or very few parental cells produced a descendent cell population. We conclude that breast cancer cells successively substitute in a tumour mass, and the high reproduction of only a portion of cancer cells may confer high adaptability to this cancer.

CCB32. Harald Vöhringer and Moritz Gerstung
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK

TensorSignatures: a multidimensional tensor factorization framework for extraction of mutational signatures

See abstract p. 32.

CCB33. Nadezda Volkova, Bettina Meier, Peter Campbell, Anton Gartner and Moritz Gerstung
European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge, UK

Mutational signatures of DNA repair deficiencies and cytotoxin exposures in *C. elegans*

Cancer is caused by alterations in the genome, which result from a combination of environmental factors damaging DNA and deficiencies in DNA repair and replication. The mutational portfolio in each individual tumour is the result of a multitude of mutagenic processes that shaped its genome over time, and the deconvolution of the observed mutational spectrum into so-called mutational signatures has attracted much attention over the past years [1]. More than 30 different signatures have been described so far, however, the causal link with the underlying mutational and repair processes is often still unclear.

In this study we used *C. elegans* as a model organism to present a systematic screen combining 25 genotoxins in different dosages under 95 different genetic conditions including single and double knock-outs of DNA repair associated genes. Upon exposure over several generations we used whole-genome sequencing to study patterns of DNA damage. We studied the mutational spectra comprising different types of genetic lesions including point mutations, indels and structural variants. We estimated the individual contributions of mutagen exposures, DNA repair deficiencies and interactions thereof using additive Poisson models. Of note, gene-gene and gene-mutagen interactions contributed nearly 30% of the explained variance.

The implication of gene-mutagen interactions is that the same damaging agent may yield different signatures in different backgrounds: the alkylating agent MMS, for example, shows only a moderate amount of T>A and T>C

mutations in a wildtype background; for translesion synthesis polymerase kappa (polk-1) deficient mutants, however, we observe a different substitution signature and a 100-fold increase in effect size, while in absence of translesion synthesis polymerase eta (polh-1) we find a complete lack of substitutions and a switch to medium-size deletions.

Different types of damage can also exhibit the same signature, e.g. if they are processed via the same repair mechanism. We show that temozolomide – a medication used for chemotherapy – and another alkylating agent EMS, which introduce methyl and ethyl groups at the O-6 position of guanine, lead to extremely similar mutational profiles despite them being chemically different and stemming from different mutagenic substances.

We also compared the patterns extracted from *C. elegans* experiments to human data, and showed that they can be successfully translated to human cancer data [2]. For mismatch repair deficient mutants, the pattern derived from *C. elegans* matched a MMR deficiency associated signature found in gastric cancers. The use of 1-bp insertions and deletions as a part of signature analysis suggested by the prevalence of these lesions in our MMR experiments also helped to refine the human MMR deficiency signature and separate it from other confounding mutational processes.

In summary, this analysis presents the first systematic catalogue of mutational signatures caused by genotoxins and DNA repair deficiencies and confirms the translational potential of these findings for cancer research.

References

[1] Alexandrov et al. Signatures of mutational processes in human cancer. Nature 2013, 500 (7463), 415-421.
[2] Meier, Volkova et al. Mutational signatures of DNA repair deficiency in C. elegans and human cancers. BioaRxiv 2018, doi: https://doi.org/10.1101/149153.

CCB34. Vyacheslav Tsyvina, Alex Zelikovsky and Pavel Skums
Department of Computer Science, Georgia State University, Atlanta, USA

Inference of tumor evolution history from single cell sequencing data using phylodynamics and discrete optimization

Cancer is a disease driven by the uncontrolled growth of cancer cells having series of somatic mutations acquired during the tumor evolution. Cancer cells form complex heterogeneous populations, which include multiple subpopulations constantly evolving to compete for resources, to metastasize, to escape immune system and therapy. Cancer heterogeneity has important implications for diagnostics and therapy, since resistant genomic variants could become dominant and lead to relapse in the patient [1], while presence of multiple variants with different frequencies may complicate detection of cancer lineages [2]. Thus development of algorithms for analysis of cancer cell populations should facilitate cancer early diagnosis and treatment. Single-cell sequencing technologies are now providing high-throughput data, which could be used to resolve and analyze intra-tumor heterogeneity at a level of single cells.

We propose a novel approach for inference of evolutionary history of heterogeneous cancer populations and estimation of relative replicative fitnesses of somatic mutations acquired over the course of tumor evolution. The proposed method also takes into account errors produced by single-cell sequencing. To achieve these goals, our method considers differences in observed relative frequencies of mutations using methods of phylodynamics, which proved to be highly useful in studying the evolution of viral populations [3]. We reconstruct cancer phylogenies which better fit the given evolutionary and sequencing error models using Bayesian inference of trees and model parameters. We use quasispecies model as an underlying evolutionary model, which has been shown to be accurate for mathematical modeling of tumorigenesis [4]. We also show how estimation of likelihoods for Bayesian inference of cancer phylogenies could be formulated and solved as discrete optimization problems, such as max-cost perfect matching in bipartite graph and network flow.

References

[1] Ding, Li, et al. "Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing." Nature 481.7382 (2012): 506.
[2] Gerlinger, Marco, et al. "Intratumor heterogeneity and branched evolution revealed by multiregion sequencing." New England journal of medicine 366.10 (2012): 883-892.
[3] Grenfell, Bryan T., et al. "Unifying the epidemiological and evolutionary dynamics of pathogens." science 303.5656 (2004): 327-332.
[4] Dominik, Wodarz, and Komarova Natalia. Computational biology of cancer: lecture notes and mathematical modeling. World Scientific, 2005.

CCB35. Coralie Williams, Barbara Adamczyk, Mohammad Afshar, Masood Kamali-Moghaddam, Niclas Karlsson, Mariana Guergova-Kuras, Frederique Lisacek, Stefan Mereiter, Frederic Parmentier, Karol Polom, Franco Roviello, Celso Reis and Qiujin Shen

Ariana Pharmaceuticals, France

Subgroups of gastric cancer patients characterized with an integrated large biomarker datasets using association rules

Availability of molecular data, characterizing cancer patients and their tumour, is required for improved diagnosis and prognosis of patients. Gastric cancer (GC) is the eighth most common deadly form of cancer worldwide, partly due to the lack of early diagnosis. The commitment of clinicians to provide a precision medicine approach in the diagnosis, prognosis and treatment of GC drives the need for better biological markers.

We describe a retrospective study collecting glycomic, proteomic, immnunohistochemistry, helicobacter pylori, and blood biomarker measurements from tissue and serum samples of 97 gastric cancer patients that underwent surgery in the Division of surgical oncology, Tertiary University Hospital of Siena, Siena, Italy.

In this work we developed a specific framework dedicated to the integration of multiple datasets from several heterogeneous sources and platforms. Experimental data was integrated with clinical, historical and survival information available for patients providing a large heterogeneous database of 951 variables.

This study identified subgroups of patients of clinical importance using a machine learning methodology (KEM®) that provides, through exhaustive exploration of all relationships between patients variables, an hypothesis driven approach to help interpret this broad database and thus identify actionable hypotheses. We systematically extracted all logical associations between experimental measures and clinical outcomes obtaining a knowledge base of 4547 associations identifying potential disease risk markers. This work helped validate previously known relationships but also served to generated new hypothesis from strong associations identified. The newly discovered markers could help early diagnosis and predict progression which could positively impact outcome for GC patients.

CCB36. Simone Zaccaria and Benjamin Raphael

Dep. of Computer Science, Princeton University, USA

Inference of allele and clone-specific copy-number aberrations in tumor samples

See abstract p. 5.

CCB37. Ron Zeira and Ron Shamir

School of Computer Science, Tel Aviv University, Israel

Sorting cancer karyotypes using double-cut-and- joins, duplications and deletions

See abstract p. 3.

CCB38. Morgan W. B. Kirzinger[1], Frederick S. Vizeacoumar[2,3], Franco J. Vizeacoumar[2*], and Anthony Kusalik[1*]

1. University of Saskatchewan, Department of Computer Science; 2. University of Saskatchewan, College of Medicine, Division of Oncology, Saskatchewan Cancer Agency; 3. University of Saskatchewan, College of Medicine, Department of Pathology; *corresponding authors

HYGIN: Identifying novel human synthetic lethal interactions through yeast orthologs

Background: With advances in cancer genomics, extensive research has been conducted to develop targeted cancer treatments by exploiting genetic interactions in cancer cells. One such genetic interaction, synthetic lethality (SL), occurs between gene pairs and is the phenomenon where the inactivation on each gene individually has no effect on cell survival, but their co-inactivation results in cell death. Though several studies have worked towards exploiting synthetic lethal interactions to develop patient specific therapeutic options, *in vitro* and *in vivo* experiments are time and resource intensive and focus primarily on known cancer associated genes. Therefore, new computational research methods are being developed to reduce the sample space and identify potential SL interactions worth exploring *in vitro*. Previously, computational studies have used a cancer directed, biased approach to identify human genetic interactions using yeast homologs of cancer genes in order to identify potential targets. These studies examined human cell line, genetic interaction, and gene mutation data sets in an attempt to identify novel drug-able targets that have SL genes. Here we develop a cancer independent human interaction network using only previously experimentally validated interactions in yeast.

Objective: Our hypothesis is that by generating a humanized yeast genetic interaction network (HYGIN) based on experimentally validated yeast orthologs, we will be able to identify potential novel patient specific gene targets for personalized medicine.

Results: We used strict one-to-one ortholog mapping for yeast and human genes to identify the SL interactions in yeast that are conserved in humans. This unbiased approach generated a humanized yeast genetic interaction network that contains 1,009 human genes and 10,419 potential SL interactions between these human genes. The network was validated using the SynLethDB, a database of known SL interactions in humans, and previous literature to show that there are interactions in common with previous *in vitro* and in silico approaches. Breast cancer gene expression data was then applied to the network to generate a breast cancer specific sub-network to identify a subset of SL interactions that are breast cancer specific. A total of 15 genes that are significantly down-regulated in breast cancer at a 2-fold cut off have a total of 115 SL partners that can be further explored for drug interaction. Our data has now generated a new interaction network that is specific to breast cancer that brings several testable hypothesis for the scientific community.

CCB39. Rebecca Sarto Basso[1], Dorit Hochbaum[1] and Fabio Vandin[2]

1. University of California, Berkeley; 2. University of Padova

Efficient algorithms to discover alterations with complementary functional association in cancer

Recent large cancer studies have measured somatic alterations in an unprecedented number of tumours. These large datasets allow the identification of cancer-related sets of genetic alterations by identifying relevant

combinatorial patterns. Among such patterns, mutual exclusivity has been employed by several recent methods that have shown its effectivenes in characterizing gene sets associated to cancer. The availability of quantitative target profiles from clinical phenotypes provides additional information that can be leveraged to improve the identification of cancer related gene sets by discovering groups with complementary functional associations with such targets. In this work we study the problem of finding groups of mutually exclusive alterations associated with a quantitative target. We propose a combinatorial formulation for the problem, and prove that the associated computation problem is NP-hard. We design two efficient algorithms, a greedy algorithm and an ILP-based algorithm, to solve the problem. We provide analytic evidence of the effectiveness of the greedy algorithm in finding high-quality solutions and show experimentally that our algorithms identify sets of alterations significantly associated with functional targets in a variety of scenarios. We show that our algorithms find sets which are better than the ones obtained by the state-of-the-art method. In addition, our algorithms are much faster than the state-of-the-art, allowing the analysis of large datasets of thousands of target profiles from cancer cell lines. We show that on one such dataset our methods identify several significant gene sets with complementary functional associations with targets.

CCB40. Urszula Czerwinska[1], Laura Cantini[1], Ulykbek Kairov[2], Emmanuel Barillot[1], Andrei Zinovyev[1]

1. Institut Curie, INSERM U900, PSL Research University, Mines ParisTech 26 rue d'Ulm, Paris; 2. Laboratory of bioinformatics and computational systems biology Center for Life Sciences, National Laboratory Astana, Nazarbayev University Astana, Kazakhstan

Application of independent component analysis to tumor transcriptomes reveals specific and reproducible immune-related signals

TME includes tumor cells, fibroblasts, and a diversity of immune cells. In many fields of science (biology, technology, sociology) observations on a studied system represent complex mixtures of signals of various origins. Tumors are engulfed in a complex microenvironment (TME) that critically impacts progression and response to therapy. In the light of recent findings, many cancer biologists believe that the state of tumor microenvironment (in particular, composition of immune system-related cells) defines the long-term effect of the cancer treatment.

Independent Component Analysis (ICA) is an unsupervised algorithm that can be used to model gene expression data as an action of a set of statistically independent hidden factors. The ICA analysis with a downstream component analysis was successfully applied to transcriptomic data previously in order to decompose bulk transcriptomic data into interpretable hidden factors. Some of these factors reflect the presence of an immune infiltrate in the tumor environment. However, no foremost studies focused on reproducibility of the ICA-based immune-related signal in the tumor transcriptome.

In this analysis, we used a set of six independent breast cancer transcriptomic datasets (BRCATCGA, METABRIC, BRCACIT, BRCABEK, BRCAWAN and BRCABCR to evaluate a detectability and a reproducibility of the immune cell-type related signal.

After ICA decomposition into 100 components, we could find major metagenes of Biton et al. representing major factors influencing tumor transcriptomes. If we focus on the IMMUNE metagene, we can find several components correlated to it. If we extract all metagenes correlated at least 0.1 (pearson correlation coefficient) with the IMMUNE metagene, we can observe clearly three clusters. Those signals were interpreted with Fisher exact test in order to evaluate enrichment in immune cell-type signatures provided from Immgen. Interestingly, we can name these three signals as T-cells, B-cells and myeloid cells with a significant p-value (< 0.01).

Our overdecomposition of six breast cancer datasets, where different normalization methods and different transcriptome profiling platforms were used, showed that the ICA-based analysis can be reproducible between datasets. It also revealed signatures that can be further used to estimate quantity of the identified immune cells in tumor transcriptomes or as a peredictive biomarkers.
Presently are working on formalizing our pipeline in a user-friendly R package that will perform signature discovery and immune cells quantification in transcriptomes of different cancer types.

CCB41. Eunyoung Kim[1], Sangwoo Kim[1]*

1. Severance Biomedical Science Institute,Brain Korea 21 PLUS Project for Medical Sciences, Yonsei UniversitiyCollege of Medicine, Korea

Comparative assessment of tools to filter whole exome sequencing data for PDX model

Patient-derived xenograft (PDX) model is commonly used in cancer research on human tumor tissue by transplanting tumor tissue obtained from patients into mouse with well-known genetic information. PDX model related study can be actively used to determine and predict the appropriate drug treatment using mutation information from PDX mouse model, which is considered to have the same genetic information as the patient. Therefore, PDX model has become a good research model in the fields of therapeutic drug development and precision medicine by finding drug targeted genes.
However, human tumor DNA in the PDX model are often captured and sequenced together with mouse tissues rather than sequenced only pure human tissue in the process of acquiring human tumor tissue from mouse. In the next-generation sequencing (NGS) analysis of PDX model, a filter process for mouse reads is essential to reduce false positive variants.
We used mouse read filter tools, Xenome, XenofilteR and Bamcmp, to test the performance to filter mouse reads in PDX sequencing data. For this in silico analysis, we used various ratio mouse read contaminated data (5%, 10%, 20%) from two human tumor whole exome sequencing (WES) data and two mouse WES data.
As a results, when applied to XenofilteR for remove mouse reads in mixture sequencing data is acquired highest sensitivity value and highest specificity and accuracy values are gained using Bamcmp tool. Comparison of performance results demonstrate Bamcmp is more appropriate tool for PDX sequencing data analysis than other tools.