

Efficient study design for the discovery of a gene expression signature predicting metastasis in cutaneous squamous cell carcinoma

Barbara Rentroia-Pacheco¹, Lara Pozza², Yan Ting Chen², Daphne Huigh²,

Celeste J. Eggermont¹, Olivia F.M. Steijlen¹, Sheril Alex², Jvalini Dwarkasing², Domenico Bellomo²,

Harmen J.G. van de Werken, Antien L. Mooyaart³, Marlies Wakkee¹, Loes M. Hollestein^{1,4}

¹Department of Dermatology, Erasmus Medical Centre, Rotterdam, The Netherlands

²SkylineDx B.V., Rotterdam, The Netherlands

³Department of Pathology, Erasmus University Medical Centre, Rotterdam, The Netherlands

⁴Department of Research, Netherlands Comprehensive Cancer Organization (IKNL), Utrecht, The Netherlands

⁵Department of Immunology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, The Netherlands

E-mail: b.rentroiapacheco@erasmusmc.nl, l.pozza@skylinedx.com, y.chen@skylinedx.com,
d.huigh@skylinedx.com, c.eggermont@erasmusmc.nl, o.steijlen@erasmusmc.nl, s.alex@skylinedx.com,
j.dwarkasing@skylinedx.com, d.bellomo@skylinedx.com, h.vandewerken@erasmusmc.nl,
a.mooyaart@erasmusmc.nl, m.wakkee@erasmusmc.nl, l.hollestein@erasmusmc.nl.

Category: PS01-23 Statistical and epidemiological methodology

Background: The Stepldent study aims to develop a gene signature predicting metastasis in patients with cutaneous squamous cell carcinoma (cSCC) to improve risk stratification, thus enabling personalized decisions about follow-up schedules and treatment options. Here we describe the unique characteristics, challenges, and best practices for an efficient design of a discovery cohort for a rare outcome (metastasis prevalence: 2-5%); for retrieving, curating, and linking the clinical and pathological data through nationwide databases; and for measuring gene expression through sequencing of archived Formalin-Fixed Paraffin-Embedded (FFPE) primary tumor samples.

Methods: Following a predefined protocol, we identified a nested-case control cohort (NCC) of 305 cases and 305 controls from a nationwide cohort of 19,120 patients with a first cSCC in the Netherlands from 2007 to 2009, followed up until 2020. We chose an NCC design since it is an efficient study design in a rare outcome setting (weighting is needed to accommodate the under-sampling of the controls). Patients were identified from the Dutch National Cancer Registry (NCR) and the clinical information was retrieved from the NCR which is linked to the nationwide registry of histo- and cytopathology (PALGA). Tumor blocks were requested from PALGA, and pathological characteristics were assessed by dermatopathologists. We matched controls to cases, based on a risk score estimated by a clinicopathological model. Gene expression was measured using the Illumina RNA Prep with Enrichment kit combined with the whole exome panel and paired-end sequenced on the NextSeq 550.

Results: Tissue slides for 541 samples were retrieved for sequencing. 151 samples were excluded after pathology review or due to low pre-library concentration. The final cohort includes 195 case-control pairs (n=390). The median sequencing depth was 43M (Q1-Q3: 35-52M); the median Q30 was 85% (Q1-Q3: 83-87%); the median GC content was 51% (Q1-Q3: 50-52%); a median of 1.8% of base pairs (Q1-Q3: 1.4-2.1%) was trimmed prior to the mapping/alignment; a median of 69% (Q1-Q3: 65-74%) of reads were aligned as protein-coding and a median of 7% (Q1-Q3: 6-10%) as rRNA; a median of 95% (Q1-Q3: 93-96%) of reads were aligned by STAR. Two samples were excluded based on quality control.

Conclusion: We described an efficient design and implementation of a nationwide discovery study in cSCC, involving the retrieval of clinicopathological data, the collection of FFPE materials, and the execution of omics measurements. This study presents the largest cohort to date, incorporating

omics measurements of primary cSCC samples, combined with simultaneous access to well-curated clinical and pathological information and follow-up data. Our findings can provide guidance for similar studies involving a rare clinical endpoint, where an efficient study design is a necessity.