Featured publications in Charles Wang Lab/Center for Genomics.

[Selected publications]
[Full list of publications].

Precision Clinical Medicine, 2023, 6: pbad004
DOI: 10.1093/pcmedi/pbad004
Research Article

Dysregulated miRNAs modulate tumor microenvironment associated signaling networks in pancreatic ductal adenocarcinoma

Tiantian Liu[1,§], Zhong Chen[1,§], Wanqiu Chen[1], Ryan Evans[2], Jane Xu[1], Mark E. Reeves[3], Michael E. de Vera[2] and Charles Wang[1,4,*]

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[2]Transplant Institute, Loma Linda University, Loma Linda, CA 92350, USA
[3]Cancer Center & School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[4]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
*Correspondence: Charles Wang, chwang@llu.edu
§Tiantian Liu and Zhong Chen contributed equally.

Abstract

The desmoplastic and complex tumor microenvironment of pancreatic ductal adenocarcinoma (PDAC) has presented tremendous challenges for developing effective therapeutic strategies. Strategies targeting tumor stroma, albeit with great potential, have met with limited success due to the lack of knowledge on the molecular dynamics within the tumor microenvironment (TME). In pursuit of a better understanding of the influence of miRNAs on TME reprogramming and to explore circulating miRNAs as diagnostic and prognostic biomarkers for PDAC, using RNA-seq, miRNA-seq, and single-cell RNA-seq (scRNA-seq), we investigated the dysregulated signaling pathways in PDAC TME modulated by miRNAs from plasma and tumor tissue. Our bulk RNA-seq in PDAC tumor tissue identified 1445 significantly differentially expressed genes with extracellular matrix and structure organization as the top enriched pathways. Our miRNA-seq identified 322 and 49 abnormally expressed miRNAs in PDAC patient plasma and tumor tissue, respectively. We found many of the TME signaling pathways were targeted by those dysregulated miRNAs in PDAC plasma. Combined with scRNA-seq from patient PDAC tumor, our results revealed that these dysregulated miRNAs were closely associated with extracellular matrix (ECM) remodeling, cell-ECM communication, epithelial-mesenchymal transition, as well as immunosuppression orchestrated by different cellular components of TME. The findings of this study could assist the development of miRNA-based stromal targeting biomarkers or therapy for PDAC patients.

Keywords: pancreatic cancer, RNA-seq, miRNA-seq, tumor microenvironment, single-cell RNA-sequencing

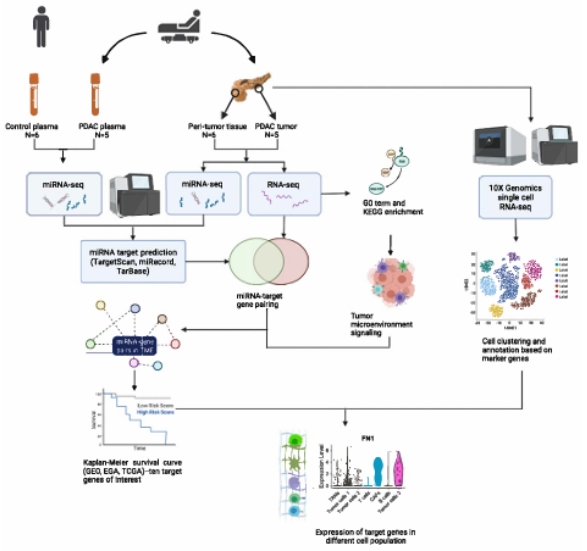


Figure 1. Overall study design.

Liu T, Chen Z, Chen W, Evans R, Xu J, Reeves ME, de Vera ME, **Wang C**. Dysregulated miRNAs modulate tumor microenvironment associated signaling networks in pancreatic ductal adenocarcinoma. doi: 10.1093/pcmedi/pbad004. PMID: 37007745; PMCID: PMC10052370.

Precision Clinical Medicine 2023 Mar 10;6(1):pbad004

communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-04737-x OPEN

Reprogramming of human peripheral blood mononuclear cells into induced mesenchymal stromal cells using non-integrating vectors

Wanqiu Chen[1], Chenguang Wang[1,2], Zhi-Xue Yang[3,4], Feng Zhang[3,4], Wei Wen[3,4], Christoph Schaniel[5], Xianqiang Mi[2], Matthew Bock[6], Xiao-Bing Zhang[3✉], Hongyu Qiu[7✉] & Charles Wang[1,8✉]

Mesenchymal stromal cells (MSCs) have great value in cell therapies. The MSC therapies have many challenges due to its inconsistent potency and limited quantity. Here, we report a strategy to generate induced MSCs (iMSCs) by directly reprogramming human peripheral blood mononuclear cells (PBMCs) with OCT4, SOX9, MYC, KLF4, and BCL-XL using a nonintegrating episomal vector system. While OCT4 was not required to reprogram PBMCs into iMSCs, omission of OCT4 significantly impaired iMSC functionality. The omission of OCT4 resulted in significantly downregulating MSC lineage specific and mesoderm-regulating genes, including *SRPX*, *COL5A1*, *SOX4*, *SALL4*, *TWIST1*. When reprogramming PBMCs in the absence of OCT4, 67 genes were significantly hypermethylated with reduced transcriptional expression. These data indicate that transient expression of OCT4 may serve as a universal reprogramming factor by increasing chromatin accessibility and promoting demethylation. Our findings represent an approach to produce functional MSCs, and aid in identifying putative function associated MSC markers.

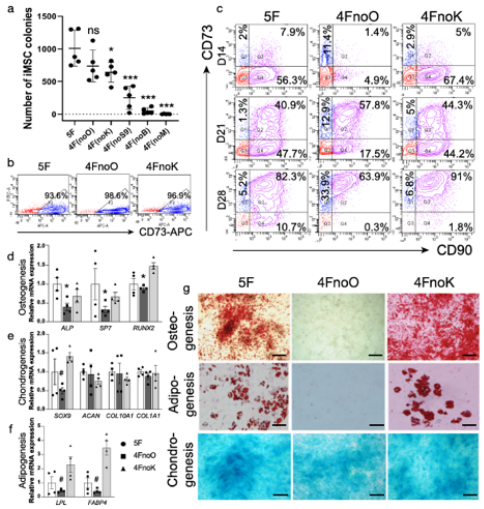


Fig. 3 Impaired trilineage differentiation potential when reprogramming without OCT4. **a** Reprogramming efficiency with the five-factor combination and removing one of the five factors. One-way ANOVA and Dunnett's multiple comparisons test, *$P < 0.05$ vs. 5F group, ***$P < 0.001$ vs. 5F group. ns: not significant. Error bars indicate standard deviation. $n = 5$ for each group from biological independent donors. **b** Flow cytometry analysis of the MSC marker CD73 4 weeks after transfection with 5F, 4FnoO (no OCT4), and 4FnoK (no KLF4). **c** Flow cytometry analysis of the MSC markers CD73 and CD90 at 2, 3, and 4 weeks after transfection with 5F, 4FnoO (no OCT4), or 4FnoK (no KLF4). **d–f** RT-qPCR analysis of osteogenesis-, adipogenesis-, and chondrogenesis-related genes in iMSCs reprogrammed with 5F, 4FnoO, and 4FnoK 2 weeks after multilineage differentiation. Tukey's multiple comparisons test, *$P < 0.05$, 4FnoO vs. 5 F and 4FnoK group. #$P < 0.05$, 4FnoO vs. 4FnoK group. $n = 4$ biologically independent samples for each group. Error bars indicate standard deviation (SD). **g** Multilineage differentiation of iMSCs reprogrammed with 5F, 4FnoO, or 4FnoK. Cells were cultured in osteogenic, adipogenic, or chondrogenic induction medium for 2–4 weeks and stained with Alizarin Red (osteogenesis), Oil Red O (adipogenesis), or Alcian blue (chondrogenesis), respectively. Scale bars represent 200 μm.

Chen W, Wang C, Yang ZX, Zhang F, Wen W, Schaniel C, Mi X, Bock M, Zhang XB, Qiu H, **Wang C** (**corresponding**). Reprogramming of human peripheral blood mononuclear cells into induced mesenchymal stromal cells using non-integrating vectors. doi: 10.1038/s42003-023-04737-x. PMID: 37041280; PMCID: PMC10090171.

Communications Biology 2023 Apr 11;6(1):393

iScience

CellPress
OPEN ACCESS

Article

Single-nucleus chromatin accessibility and RNA sequencing reveal impaired brain development in prenatally e-cigarette exposed neonatal rats

Zhong Chen,[1,5] Wanqiu Chen,[1,5] Yong Li,[2,5] Malcolm Moos, Jr.,[3] Daliao Xiao,[2,*] and Charles Wang[1,4,6,*]

SUMMARY

Although emerging evidence reveals that vaping alters the function of the central nervous system, the effects of maternal vaping on offspring brain development remain elusive. Using a well-established *in utero* exposure model, we performed single-nucleus ATAC-seq (snATAC-seq) and RNA sequencing (snRNA-seq) on prenatally e-cigarette-exposed rat brains. We found that maternal vaping distorted neuronal lineage differentiation in the neonatal brain by promoting excitatory neurons and inhibiting lateral ganglionic eminence-derived inhibitory neuronal differentiation. Moreover, maternal vaping disrupted calcium homeostasis, induced microglia cell death, and elevated susceptibility to cerebral ischemic injury in the developing brain of offspring. Our results suggest that the aberrant calcium signaling, diminished microglial population, and impaired microglia-neuron interaction may all contribute to the underlying mechanisms by which prenatal e-cigarette exposure impairs neonatal rat brain development. Our findings raise the concern that maternal vaping may cause adverse long-term brain damage to the offspring.

Chen Z, Chen W, Li Y, Moos M Jr, Xiao D, **Wang C (corresponding)**. Single-nucleus chromatin accessibility and RNA sequencing reveal impaired brain development in prenatally e-cigarette exposed neonatal rats. doi: 10.1016/j.isci.2022.104686. PMID: 35874099; PMCID: PMC9304611.

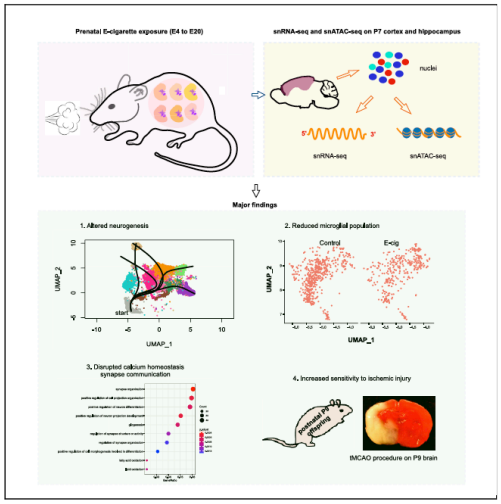
iScience 2022 Jun 30;25(8):104686

iScience

CellPress
OPEN ACCESS

Article

Single-nucleus chromatin accessibility and RNA sequencing reveal impaired brain development in prenatally e-cigarette exposed neonatal rats



Zhong Chen, Wanqiu Chen, Yong Li, Malcolm Moos, Jr., Daliao Xiao, Charles Wang

dxiao@lu.edu (D.X.)
chwang@lu.edu (C.W.)

Highlights

Prenatal e-cigarette exposure adversely affects rat brain development

Prenatal e-cigarette smoking disrupts brain excitatory/inhibitory neuron balance

Prenatal e-cigarette smoking disrupts Ca^{2+} homeostasis, induces microglia cell death

Prenatal e-cigarette exposure increases susceptibility to brain ischemic injury

Chen et al., iScience 25, 104686
August 19, 2022
https://doi.org/10.1016/j.isci.2022.104686

iScience
Article

CellPress
OPEN ACCESS

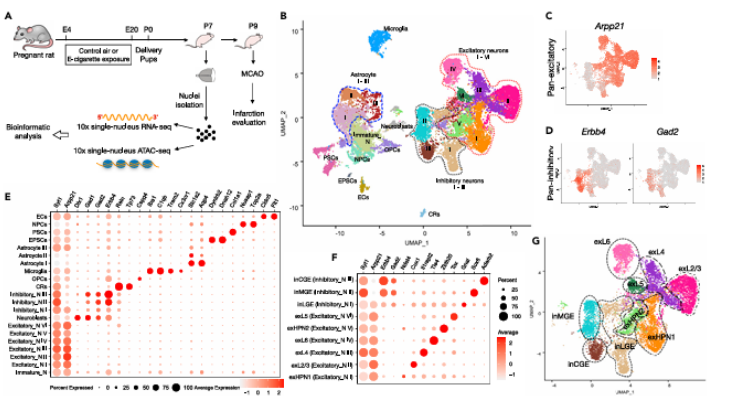


Figure 1. Study design and cell type identification in the neonatal rat brain exposed to prenatal e-cigarette vapor

(A) Overview of study design. Pregnant Sprague-Dawley rats were exposed to control air or e-cigarette from gestational E4 to E20. The brains of postnatal day 7 (P7) rat pups were dissociated to obtain single nuclei for snRNA-seq and snATAC-seq. In some parallel animals, transient middle cerebral artery occlusion (tMCAO) was conducted on postnatal day 9 (P9) rat pups, and the infarct size was evaluated 48h later.

(B) UMAP visualization of 29,317 single nucleus transcriptomes from postnatal day 7 rat brain revealed 21 groups colored by annotated cell type. OPCs, oligodendrocyte progenitor cells; NPCs, neural progenitor cells; PSCs, perivascular stromal cells; EPSCs, ependymal cells; ECs, endothelial cells; CRs, Cajal-Retzius cells.

(C) UMAP visualization of nine neuronal cell populations showing the expression of the representative excitatory neuron marker *Arpp21* and (D). inhibitory neuronal markers *Erbb4* and *Gad2*.

(E) Dot plot showing representative marker gene expression across all 21 cell clusters derived from P7 neonatal rat brain cortex and hippocampus.

(F) Dot plot of representative marker gene expression in neuronal subpopulations.

(G) UMAP visualization of nine neuronal cell populations showing the annotation of each neuronal subtype.

GUT MICROBES
2022, VOL. 14, NO. 1, e2125733 (20 pages)
https://doi.org/10.1080/19490976.2022.2125733

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Rural environment reduces allergic inflammation by modulating the gut microbiota

Zhaowei Yang[a#], Zhong Chen[b#], Xinliu Lin[a#], Siyang Yao[a#], Mo Xian[a#], Xiaoping Ning[a#], Wanyi Fu[a#], Mei Jiang[a], Naijian Li[a], Xiaojun Xiao[c], Mulin Feng[a], Zexuan Lian[a], Wenqing Yang[a], Xia Ren[a], Zhenyu Zheng[a], Jiefeng Zhao[a], Nili Wei[a], Wenju Lu[a], Marjut Roponen[d], Bianca Schaub[e], Gary W. K. Wong[f], Zhong Su[a], Charles Wang[b], and Jing Li[a]

[a]Department of Allergy and Clinical Immunology, State Key Laboratory of Respiratory Disease, Guangzhou Institute of Respiratory Health, the First Affiliated Hospital of Guangzhou Medical University, Guangzhou, Guangdong, P. R. China; [b]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA USA; [c]State Key Laboratory of Respiratory Disease for Allergy at Shenzhen University, Shenzhen Key Laboratory of Allergy and Immunology, Shenzhen University School of Medicine, Shenzhen, China; [d]Department of Environmental Science, University of Eastern Finland, Kuopio, Finland; [e]Department of Pulmonary and Allergy, University Children's Hospital Munich, LMU Munich, Munich, Germany; [f]Department of Paediatrics, Prince of Wales Hospital, the Chinese University of Hong Kong, Hong Kong, China; [g]State Key Laboratory of Respiratory Disease, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou, China

ABSTRACT

Rural environments and microbiota are linked to a reduction in the prevalence of allergies. However, the mechanism underlying the reduced allergies modulated by rural residency is unclear. Here, we assessed gut bacterial composition and metagenomics in urban and rural children in the EuroPrevall-INCO cohort. Airborne dusts, including mattress and rural henhouse dusts, were profiled for bacterial and fungal composition by amplicon sequencing. Mice were repeatedly exposed to intranasal dust extracts and evaluated for their effects on ovalbumin (OVA)-induced allergic airway inflammation, and gut microbiota restoration was validated by fecal microbiota transplant (FMT) from dust-exposed donor mice. We found that rural children had fewer allergies and unique gut microbiota with fewer *Bacteroides* and more *Prevotella*. Indoor dusts in rural environments harbored higher endotoxin level and diversity of bacteria and fungi, whereas indoor urban dusts were enriched with *Aspergillus* and contained elevated pathogenic bacteria. Intranasal administration of rural dusts before OVA sensitization reduced respiratory eosinophils and blood IgE level in mice and also led to a recovery of gut bacterial diversity and *Ruminiclostridium* in the mouse model. FMT restored the protective effect by reducing OVA-induced lung eosinophils in recipient mice. Together, these results support a cause-effect relationship between exposure to dust microbiota and allergy susceptibility in children and mice. Specifically, rural environmental exposure modulated the gut microbiota, which was essential in reducing allergy in children from Southern China. Our findings support the notion that the modulation of gut microbiota by exposure to rural indoor dust may improve allergy prevention.

ARTICLE HISTORY
Received 21 July 2022
Revised 6 September 2022
Accepted 8 September 2022

KEYWORDS
Allergies; *Aspergillus*; gut microbiota; indoor microbiota; urban/rural environments; 16S rRNA-seq and metagenomics

Yang Z, Chen Z, Lin X, Yao S, Xian M, Ning X, Fu W, Jiang M, Li N, Xiao X, Feng M, Lian Z, Yang W, Ren X, Zheng Z, Zhao J, Wei N, Lu W, Roponen M, Schaub B, Wong GWK, Su Z, **Wang C** (**co-corresponding**), Li J. Rural environment reduces allergic inflammation by modulating the gut microbiota. doi: 10.1080/19490976.2022.2125733. PMID: 36193874; PMCID: PMC9542937. 122.

Gut Microbes 2022 Jan-Dec;14(1):2125733

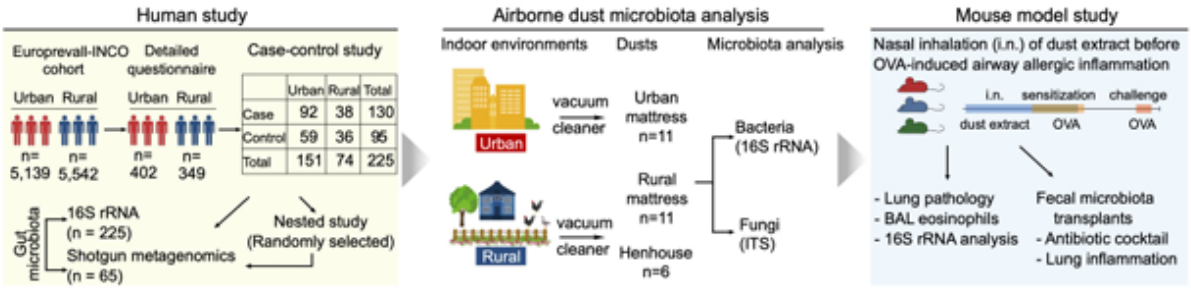


Figure 1. Overall study design. The study included (a) a human study, (b) an airborne dust microbiota analysis, and (c) a mouse model study. 16S rRNA and ITS, 16S rRNA and Internal Transcribed Spacer amplicon sequencing; BAL, bronchoalveolar lavage; EuroPrevall-INCO, a project to evaluate the prevalence of food allergy in China, India, and Russia using the standardized methodology of the EuroPrevall protocol;[18] OVA, ovalbumin. Mouse color (panel c) indicates dust extracts from urban mattresses, rural mattresses, and henhouses. Individual dust from each environment was pooled and used in the mouse model study.

Xiao et al. Genome Biology (2022) 23:237
https://doi.org/10.1186/s13059-022-02803-x

Genome Biology

RESEARCH Open Access

Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples

Chunlin Xiao[1*], Zhong Chen[2], Wanqiu Chen[2], Cory Padilla[3], Michael Colgan[4], Wenjun Wu[5], Li-Tai Fang[6], Tiantian Liu[2], Yibin Yang[5], Valerie Schneider[1], Charles Wang[2*] and Wenming Xiao[4*]

*Correspondence: xiao2@mail.nih.gov; chwang@llu.edu; Wenming.Xiao@fda.hhs.gov

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20894, USA
[2] Center for Genomics, Loma Linda University School of Medicine, 11021 Campus St., Loma Linda, CA 92350, USA
[3] Dovetail Genomics, 100 Enterprise Way, Scotts Valley, CA 95066, USA
[4] The Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, USA
[5] Blood Cell Development and Function Program, Fox Chase Cancer Center, Philadelphia, PA 19111, USA
[6] Bioinformatics Research & Early Development, Roche Sequencing Solutions Inc., 1301 Shoreway Road, Belmont, CA 94002, USA

Abstract

Background: The use of a personalized haplotype-specific genome assembly, rather than an unrelated, mosaic genome like GRCh38, as a reference for detecting the full spectrum of somatic events from cancers has long been advocated but has never been explored in tumor-normal paired samples. Here, we provide the first demonstrated use of de novo assembled personalized genome as a reference for cancer mutation detection and quantifying the effects of the reference genomes on the accuracy of somatic mutation detection.

Results: We generate de novo assemblies of the first tumor-normal paired genomes, both nuclear and mitochondrial, derived from the same individual with triple negative breast cancer. The personalized genome was chromosomal scale, haplotype phased, and annotated. We demonstrate that it provides individual specific haplotypes for complex regions and medically relevant genes. We illustrate that the personalized genome reference not only improves read alignments for both short-read and long-read sequencing data but also ameliorates the detection accuracy of somatic SNVs and SVs. We identify the equivalent somatic mutation calls between two genome references and uncover novel somatic mutations only when personalized genome assembly is used as a reference.

Conclusions: Our findings demonstrate that use of a personalized genome with individual-specific haplotypes is essential for accurate detection of the full spectrum of somatic mutations in the paired tumor-normal samples. The unique resource and methodology established in this study will be beneficial to the development of precision oncology medicine not only for breast cancer, but also for other cancers.

Xiao C, Chen Z, Chen W, Padilla C, Colgan M, Wu W, Fang LT, Liu T, Yang Y, Schneider V, **Wang C (co-corresponding)**, Xiao W. Personalized genome assembly for accurate cancer somatic mutation discovery using tumor-normal paired reference samples. doi: 10.1186/s13059-022-02803-x. PMID: 36352452; PMCID: PMC9648002.

Genome Biology 2022 Nov 9;23(1):237

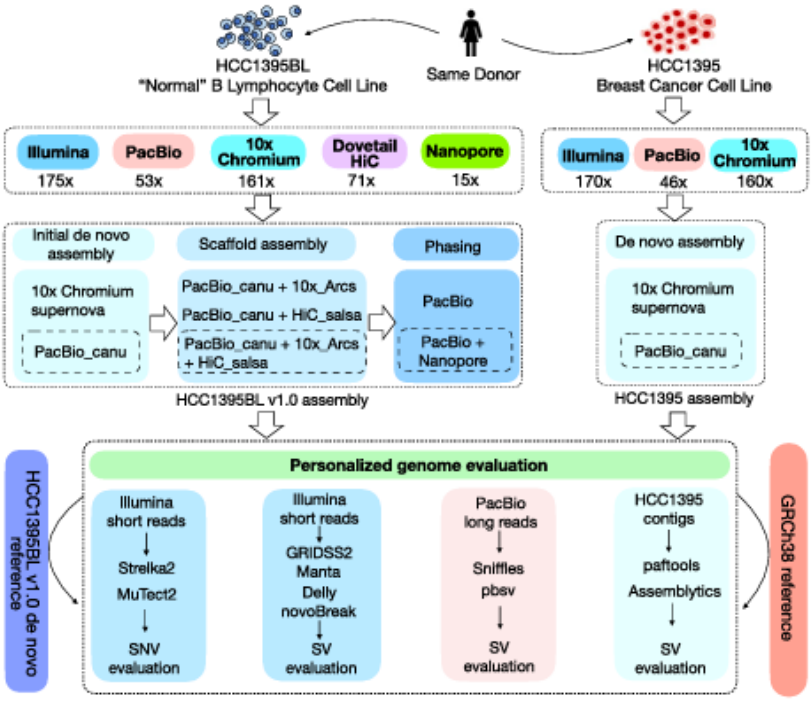


Fig. 1 Schematic diagram of study design. Sequencing data from five different platforms were used for initial de novo assembly, assembly evaluation, scaffolding, and phasing for the normal reference sample (HCC1395BL B Lymphocyte cell line), while sequencing data from three platforms were used for de novo assembly and assembly evaluation for the tumor reference sample (HCC1395 breast cancer cell line from the same donor). The final assembled personal genome, known as HCC1395BL_v1.0, was used as reference for read mapping with both short and long reads, and assessment of somatic SNVs and SVs as compared to that

RESEARCH Open Access

Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies

Keyur Talsania[1,2†], Tsai-wei Shen[1,2†], Xiongfong Chen[1,2†], Erich Jaeger[3†], Zhipan Li[4†], Zhong Chen[5], Wanqiu Chen[5], Bao Tran[6], Rebecca Kusko[7], Limin Wang[8], Andy Wing Chun Pang[9], Zhaowei Yang[10], Sulbha Choudhari[1,2], Michael Colgan[11], Li Tai Fang[12], Andrew Carroll[13], Jyoti Shetty[6], Yuliya Kriga[6], Oksana German[6], Tatyana Smirnova[6], Tiantain Liu[5], Jing Li[10], Ben Kellman[9], Karl Hong[9], Alex R. Hastie[9], Aparna Natarajan[3], Ali Moshrefi[3], Anastasiya Granat[3], Tiffany Truong[3], Robin Bombardi[3], Veronnica Mankinen[14], Daoud Meerzaman[15], Christopher E. Mason[16], Jack Collins[1,2], Eric Stahlberg[2], Chunlin Xiao[17], Charles Wang[5*], Wenming Xiao[11*] and Yongmei Zhao[1,2*]

†Keyur Talsania, Tsai-wei Shen, Xiongfong Chen, Erich Jaeger, Zhipan Li contributed equally.

*Correspondence: chwang@llu.edu; Wenming.Xiao@fda.hhs.gov; Yongmei.Zhao@nih.gov

[2] Bioinformatics and Computational Science Directorate, Frederick National Laboratory for Cancer Research, Frederick, MD, USA
[5] Center for Genomics, Loma Linda University School of Medicine, Loma Linda, CA, USA
[11] Center for Drug Evaluation and Research, FDA, Silver Spring, MD, USA
Full list of author information is available at the end of the article

Abstract

Background: The cancer genome is commonly altered with thousands of structural rearrangements including insertions, deletions, translocation, inversions, duplications, and copy number variations. Thus, structural variant (SV) characterization plays a paramount role in cancer target identification, oncology diagnostics, and personalized medicine. As part of the SEQC2 Consortium effort, the present study established and evaluated a consensus SV call set using a breast cancer reference cell line and matched normal control derived from the same donor, which were used in our companion benchmarking studies as reference samples.

Results: We systematically investigated somatic SVs in the reference cancer cell line by comparing to a matched normal cell line using multiple NGS platforms including Illumina short-read, 10X Genomics linked reads, PacBio long reads, Oxford Nanopore long reads, and high-throughput chromosome conformation capture (Hi-C). We established a consensus SV call set of a total of 1788 SVs including 717 deletions, 230 duplications, 551 insertions, 133 inversions, 146 translocations, and 11 breakends for the reference cancer cell line. To independently evaluate and cross-validate the accuracy of our consensus SV call set, we used orthogonal methods including PCR-based validation, Affymetrix arrays, Bionano optical mapping, and identification of fusion genes detected from RNA-seq. We evaluated the strengths and weaknesses of each NGS technology for SV determination, and our findings provide an actionable guide to improve cancer genome SV detection sensitivity and accuracy.

Talsania K, Shen TW, Chen X, Jaeger E, Li Z, Chen Z, Chen W, Tran B, Kusko R, Wang L, Pang AWC, Yang Z, Choudhari S, Colgan M, Fang LT, Carroll A, Shetty J, Kriga Y, German O, Smirnova T, Liu T, Li J, Kellman B, Hong K, Hastie AR, Natarajan A, Moshrefi A, Granat A, Truong T, Bombardi R, Mankinen V, Meerzaman D, Mason CE, Collins J, Stahlberg E, Xiao C, **Wang C (co-corresponding),** Xiao W, Zhao Y. Structural variant analysis of a cancer reference cell line sample using multiple sequencing technologies. doi: 10.1186/s13059-022-02816-6. PMID: 36514120; PMCID: PMC9746098.

Genome Biology 2022 Dec 13;23(1):255

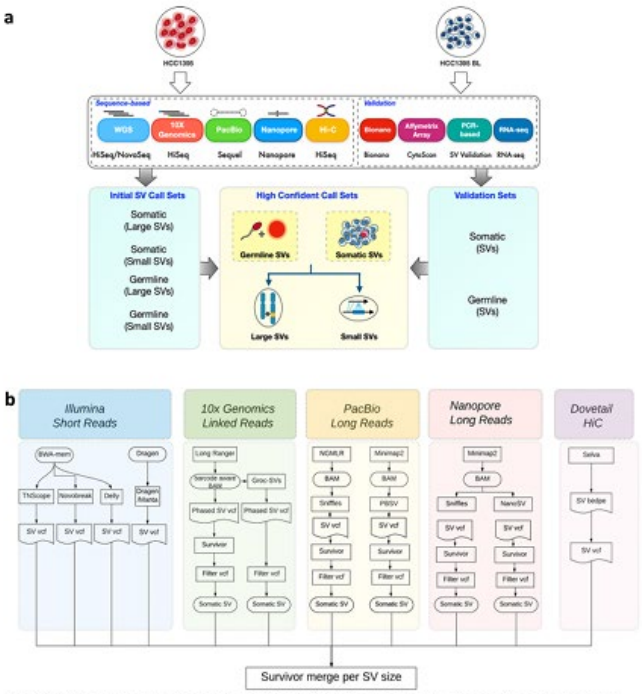


Fig. 1 Study design and bioinformatics workflow for SV detection and integration. **a** Schematic overview of the study design. Two well-characterized reference cell lines (HCC1395 and HCC1395BL) were used to generate whole-genome sequencing data across five platforms (Illumina short reads, 10X Genomics linked reads, PacBio long reads, Oxford Nanopore long reads, and Dovetail Hi-C proximity ligation). Initial SV call sets were identified from each platform and combined together to identify high-confidence call sets. The SVs from high-confidence call sets were selected for PCR-based validation for deletion, insertion, intra-chromosomal inversion and inter-chromosomal translocation; copy number changes were validated by Affymetrix array. Large SVs (≥20kb) were validated using Bionano optical mapping. RNA-seq was used to validate the fusion gene and translocation events. **b** Schematic overview of the bioinformatics analysis workflow. Each platform's data was processed by the aligner and SV caller specific to that platform. The tumor-only or somatic SV calls were selected by Survivor. The final call sets from each platform were integrated together using the Survivor software tool based a window-size approach and SV types

OXFORD

Precision Clinical Medicine, 4(1), 2021, 1–16

doi: 10.1093/pcmedi/pbab001
Advance Access Publication Date: 18 January 2021
Research Article

RESEARCH ARTICLE

Computational molecular docking and virtual screening revealed promising SARS-CoV-2 drugs

Maryam Hosseini[1], Wanqiu Chen[1], Daliao Xiao[2] and Charles Wang[1,3,*]

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[2]Lawrence D. Longo, MD Center for Perinatal Biology, Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[3]Division of Microbiology & Molecular Genetics, Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA

*Correspondence: Charles Wang, chwang@llu.edu or oxwang@gmail.com
Daliao Xiao, http://orcid.org/http://orcid.org/0000-0003-0147-2758
Charles Wang, http://orcid.org/http://orcid.org/0000-0001-8861-2121

Abstract

The pandemic of novel coronavirus disease 2019 (COVID-19) has rampaged the world, with more than 58.4 million confirmed cases and over 1.38 million deaths across the world by 23 November 2020. There is an urgent need to identify effective drugs and vaccines to fight against the virus. Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) belongs to the family of coronaviruses consisting of four structural and 16 non-structural proteins (NSP). Three non-structural proteins, main protease (Mpro), papain-like protease (PLpro), and RNA-dependent RNA polymerase (RdRp), are believed to have a crucial role in replication of the virus. We applied computational ligand-receptor binding modeling and performed comprehensive virtual screening on FDA-approved drugs against these three SARS-CoV-2 proteins using AutoDock Vina, Glide, and rDock. Our computational studies identified six novel ligands as potential inhibitors against SARS-CoV-2, including antiemetics rolapitant and ondansetron for Mpro; labetalol and levomefolic acid for PLpro; and leucal and antifungal natamycin for RdRp. Molecular dynamics simulation confirmed the stability of the ligand-protein complexes. The results of our analysis with some other suggested drugs indicated that chloroquine and hydroxychloroquine had high binding energy (low inhibitory effect) with all three proteins—Mpro, PLpro, and RdRp. In summary, our computational molecular docking approach and virtual screening identified some promising candidate SARS-CoV-2 inhibitors that may be considered for further clinical studies.

Key words: COVID-19; SARS-CoV-2; Mpro; PLpro; RdRp; virtual screening; molecular docking

Hosseini M, Chen W, Xiao D, **Wang C (corresponding)**. Computational molecular docking and virtual screening revealed promising SARS-CoV-2 drugs. doi: 10.1093/pcmedi/pbab001. PMID: 33842834; PMCID: PMC7928605.

Precision Clinical Medicine 2021 Jan 18;4(1):1-16

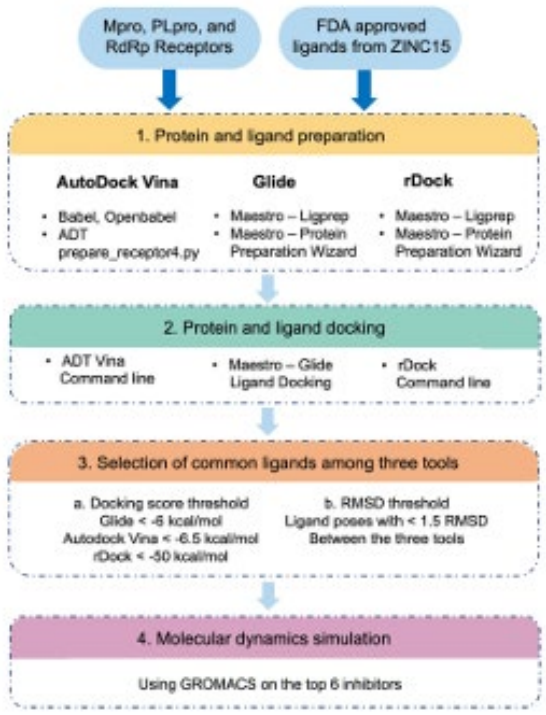


Figure 1. General scheme of the docking protocol.

docking scores. The docking processes were done using written in-house scripts. All visualizations were done using Schrodinger Maestro[37] and PyMOL (https://pymol.org/).

OXFORD

Precision Clinical Medicine, 4(2), 2021, 93–108

10.1093/pcmedi/pbab012
Advance Access Publication Date: 17 May 2021
Research Article

RESEARCH ARTICLE

Spaceflight decelerates the epigenetic clock orchestrated with a global alteration in DNA methylome and transcriptome in the mouse retina

Zhong Chen[1,§], Seta Stanbouly[1,2,§], Nina C. Nishiyama[2], Xin Chen[1], Michael D. Delp[3], Hongyu Qiu[4], Xiao W. Mao[2] and Charles Wang[1,5,*]

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[2]Division of Radiation Research, Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92354, USA
[3]Department of Nutrition and Integrative Physiology, Florida State University, Tallahassee, FL 32306, USA
[4]Center for Molecular and Translational Medicine, Institute of Biomedical Science, Georgia State University, Atlanta, GA 30303, USA
[5]Division of Microbiology and Molecular Genetics, Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA

*Correspondence: Charles Wang, chwang@llu.edu
Charles Wang, http://orcid.org/0000-0001-8861-2121
§These authors contributed equally to this work.

Abstract

Astronauts exhibit an assortment of clinical abnormalities in their eyes during long-duration spaceflight. The purpose of this study was to determine whether spaceflight induces epigenomic and transcriptomic reprogramming in the retina or alters the epigenetic clock. The mice were flown for 37 days in animal enclosure modules on the International Space Station; ground-based control animals were maintained under similar housing conditions. Mouse retinas were isolated and both DNA methylome and transcriptome were determined by deep sequencing. We found that a large number of genes were differentially methylated with spaceflight, whereas there were fewer differentially expressed genes at the transcriptome level. Several biological pathways involved in retinal diseases such as macular degeneration were significantly altered. Our results indicated that spaceflight decelerated the retinal epigenetic clock. This study demonstrates that spaceflight impacts the retina at the epigenomic and transcriptomic levels, and such changes could be involved in the etiology of eye-related disorders among astronauts.

Key words: spaceflight; DNA methylome; transcriptome; epigenetic clock

Chen Z, Stanbouly S, Nishiyama NC, Chen X, Delp MD, Qiu H, Mao XW, **Wang C (corresponding)**. Spaceflight decelerates the epigenetic clock orchestrated with a global alteration in DNA methylome and transcriptome in the mouse retina. doi: 10.1093/pcmedi/pbab012. PMID: 34179686; PMCID: PMC8220224.

Precision Clinical Medicine 2021 May 17;4(2):93-108

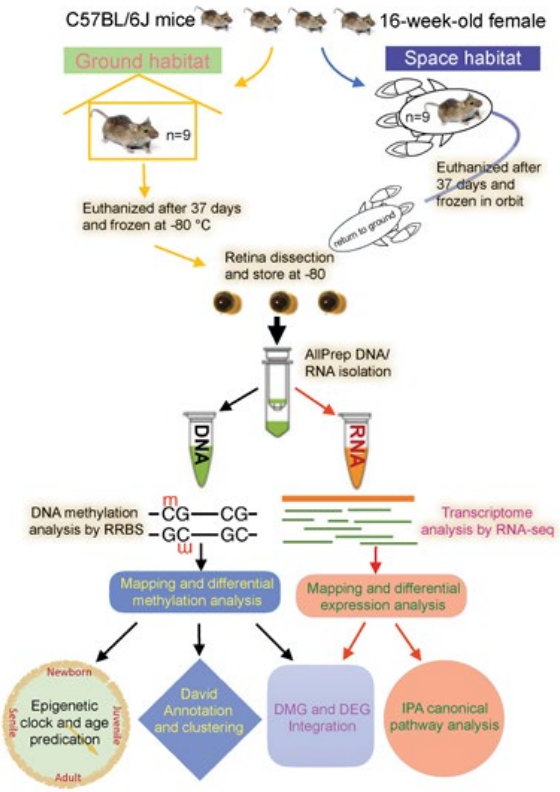


Figure 1. Study design. This study was under the NASA Rodent Research-1 (RR-1) project consortium. Spaceflight mice after flight were euthanized and frozen on orbit. The retinas were isolated from both the ground habitat control and flight mice after returning to ground. RNA and DNA were extracted from the retinas, and then RRBS and RNA-seq libraries were constructed to obtain DNA methylome and transcriptome, respectively. Retinal epigenetic age was calculated using MouseEpigeneticClock script. Disease and bio-functional pathways were generated using the DAVID GO functional annotation and the Ingenuity Pathway through Analysis (IPA) based on DMGs & DEGs.

iScience

CellPress
OPEN ACCESS

Article

A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples

Tiantian Liu,[1,7] Zhong Chen,[1,7] Wanqiu Chen,[1,7] Xin Chen,[1,2] Maryam Hosseini,[1] Zhaowei Yang,[1,3] Jing Li,[2,3] Diana Ho,[1] David Turay,[4] Ciprian P. Gheorghe,[5] Wendell Jones,[6,*] and Charles Wang[1,2,8,*]

SUMMARY

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is an emerging new type of coronavirus that is responsible for the COVID-19 pandemic and the unprecedented global health emergency. Whole-genome sequencing (WGS) of SARS-CoV-2 plays a critical role in understanding the disease. Performance variation exists across SARS-CoV-2 viral WGS technologies, but there is currently no benchmarking study comparing different WGS sequencing protocols. We compared seven different SARS-CoV-2 WGS library protocols using RNA from patient nasopharyngeal swab samples under two storage conditions with low and high viral inputs. We found large differences in mappability and genome coverage, and variations in sensitivity, reproducibility, and precision of single-nucleotide variant calling across different protocols. For certain amplicon-based protocols, an appropriate primer trimming step is critical for accurate single-nucleotide variant calling. We ranked the performance of protocols based on six different metrics. Our findings offer guidance in choosing appropriate WGS protocols to characterize SARS-CoV-2 and its evolution.

INTRODUCTION

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA, USA

[2]Division of Microbiology & Molecular Genetics, Department of Basic Science, School of Medicine, Loma Linda University, Loma Linda, CA, USA

Liu T, Chen Z, Chen W, Chen X, Hosseini M, Yang Z, Li J, Ho D, Turay D, Gheorghe CP, Jones W, **Wang C (corresponding)**. A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples. doi: 10.1016/j.isci.2021.102892. Epub 2021 Jul 21. PMID: 34308277; PMCID: PMC8294598.

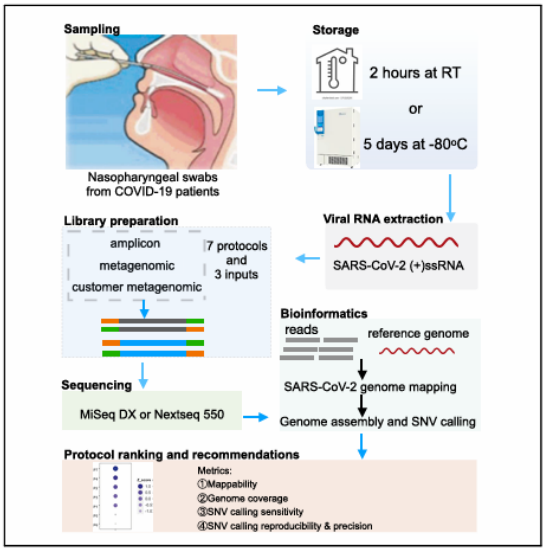
iScience 2021 Aug 20;24(8):102892

iScience

CellPress
OPEN ACCESS

Article

A benchmarking study of SARS-CoV-2 whole-genome sequencing protocols using COVID-19 patient samples



Tiantian Liu, Zhong Chen, Wanqiu Chen, ..., Ciprian P. Gheorghe, Wendell Jones, Charles Wang

wendell.jones@q2labsolutions.com (W.J.)
oxwang@gmail.com, chwang@llu.edu (C.W.)

Highlights

Benchmarking comparison on seven different SARS-CoV-2 genome sequencing protocols

Ranked protocol performances based on genome coverage and SNV detection

Raised concern of primer-masking on SNV calling from Qiagen Amplicon protocol

Proposed bioinformatics remedy on making correct SNV call from masked genome regions

iScience
Article

CellPress
OPEN ACCESS

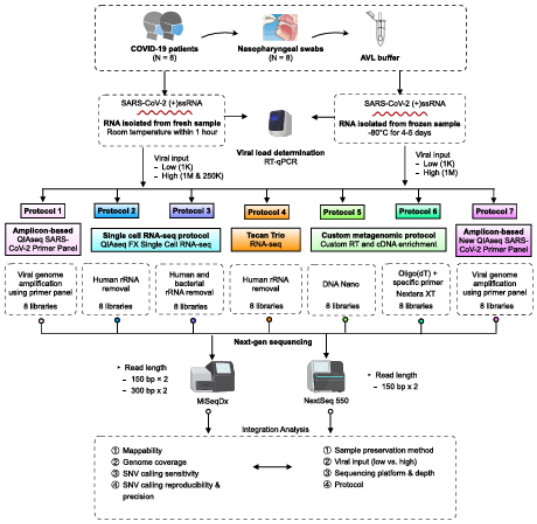


Figure 1. Schematic overview of the experimental design and workflow

Eight COVID-19 positive patient nasopharyngeal swab samples were used to construct the SARS-CoV-2 WGS libraries using seven protocols. Two different sample storage conditions were compared. For fresh samples, three different viral inputs, i.e., 1000 (1K, low) vs. either 250,000 or 1 million (250K or 1M, high) SARS-CoV-2 viral copies, were used from each same sample, whereas for frozen samples, two different viral inputs, i.e., 1000 (1K, low) vs. 1 million (1M, high) SARS-CoV-2 viral copies from each same sample, were used. P4 used different samples at low input vs. high input due to minimal total RNA amount required. The performances of protocols were benchmarked based on viral input, sequencing platform and depth, mappability, viral genome coverage and coverage uniformity, and sensitivity, reproducibility, as well as precision across seven protocols.

nature biotechnology

FOCUS | ARTICLES

https://doi.org/10.1038/s41587-020-00748-9

Check for updates

A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples

Wanqiu Chen[1,12], Yongmei Zhao[2,3,12], Xin Chen[1,4,12], Zhaowei Yang[1,5,12], Xiaojiang Xu[6], Yingtao Bi[7], Vicky Chen[2,3], Jing Li[4,5], Hannah Choi[1], Ben Ernest[8], Bao Tran[3], Monika Mehta[3], Parimal Kumar[3], Andrew Farmer[9], Alain Mir[9], Urvashi Ann Mehra[8], Jian-Liang Li[6], Malcolm Moos Jr.[10], Wenming Xiao[11] ✉ and Charles Wang[1,4] ✉

Comparing diverse single-cell RNA sequencing (scRNA-seq) datasets generated by different technologies and in different laboratories remains a major challenge. Here we address the need for guidance in choosing algorithms leading to accurate biological interpretations of varied data types acquired with different platforms. Using two well-characterized cellular reference samples (breast cancer cells and B cells), captured either separately or in mixtures, we compared different scRNA-seq platforms and several preprocessing, normalization and batch-effect correction methods at multiple centers. Although preprocessing and normalization contributed to variability in gene detection and cell classification, batch-effect correction was by far the most important factor in correctly classifying the cells. Moreover, scRNA-seq dataset characteristics (for example, sample and cellular heterogeneity and platform used) were critical in determining the optimal bioinformatic method. However, reproducibility across centers and platforms was high when appropriate bioinformatic methods were applied. Our findings offer practical guidance for optimizing platform and software selection when designing an scRNA-seq study.

Chen W, Zhao Y, Chen X, Yang Z, Xu X, Bi Y, Chen V, Li J, Choi H, Ernest B, Tran B, Mehta M, Kumar P, Farmer A, Mir A, Mehra UA, Li JL, Moos M Jr, Xiao W, **Wang C (corresponding)**. A multicenter study benchmarking single-cell RNA sequencing technologies using reference samples. doi: 10.1038/s41587-020-00748-9. Epub 2020 Dec 21. PMID: 33349700.

Nature Biotechnology 2021 Sep;39(9):1103-1114

Received the Best Paper Award, at the 2021 MAQC Society Annual Meeting!

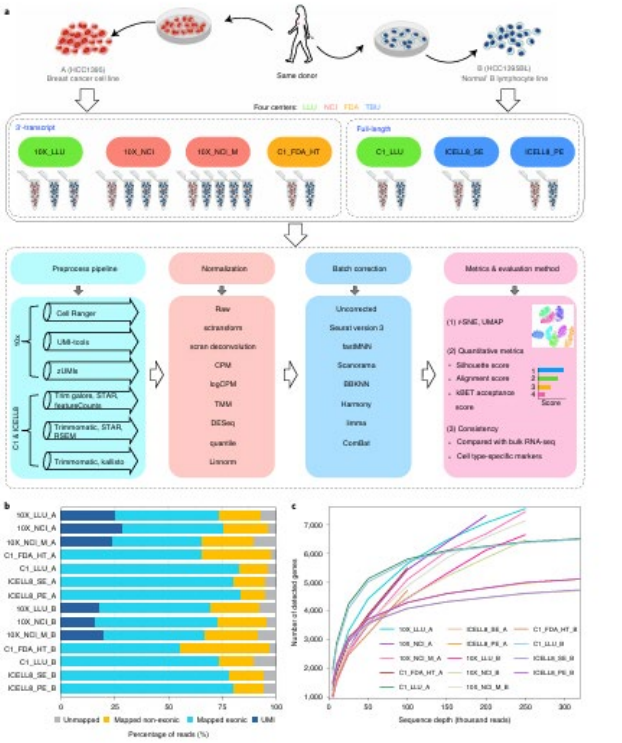


Fig. 1 | Overall study design, scRNA-seq mapping and numbers of genes detected across datasets. a, Schematic overview of the study design (see detailed descriptions and notations in the Methods). Two reference cell lines (sample A, HCC1395; sample B, HCC1395BL) were used to generate scRNA-seq data across four platforms (10x Genomics, Fluidigm C1 HT, Fluidigm C1 and Takara Bio ICELL8) and four testing sites (LLU, NCI, FDA and TBU). At the LLU and NCI sites (10x), mixed single-cell captures and library constructions were also prepared with either 10% or 5% cancer cells spiked into the B lymphocytes. At the NCI site, single-cell captures and library constructions were also performed with two methanol-fixed cell mixtures (5% cancer cells spiked into B lymphocytes, termed fixed_1 and fixed_2). One set of 10x scRNA libraries from the NCI was also sequenced using a shorter modified sequencing method. BK RNA-seq data were also obtained from these cell lines, each in triplicate. See Methods for details about study design. **b**, For both the breast cancer cell line (sample A) and the B lymphocyte line (sample B) across 14 pairwise datasets, percentages are shown of reads that mapped to the exonic region (blue) or the non-exonic region (orange) or did not map to the human genome (gray). For UMI methods (10x), dark blue indicates the exonic reads with UMIs. **c**, Median number of genes detected per cell at different sequencing read depths.

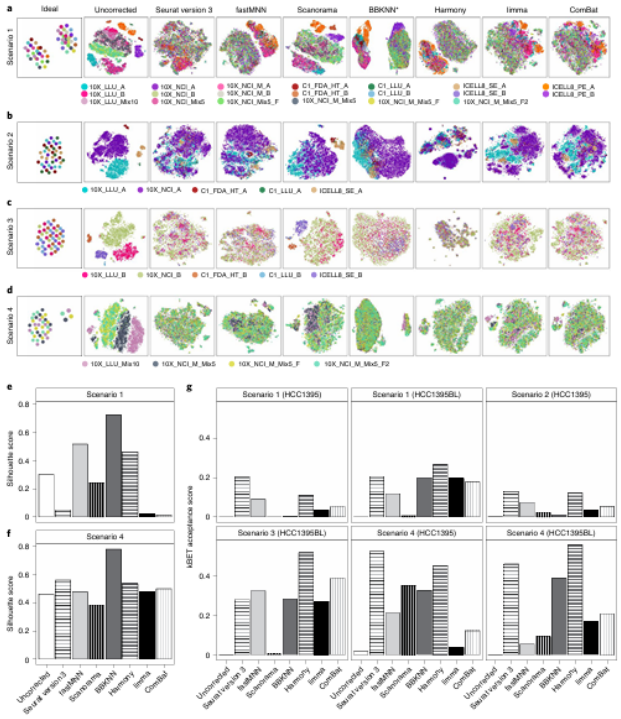


Fig. 4 | Batch-effect corrections evaluated in four different sample composition scenarios. a, Batch-effect correction in scenario 1, in which all 20 scRNA-seq datasets were combined, including mixed and unmixed datasets, with large proportions of two dissimilar types of cells (sample A, breast cancer cell line HCC1395; sample B, B lymphocyte line HCC1395BL). Datasets from the 10x platform were downsampled to 1,200 cells per dataset. **b**, Batch-effect correction in scenario 2, in which five scRNA-seq datasets (10X_LLU_A, 10X_NCI_A, C1_FDA_HT_A, C1_LLU_A and ICELL8_SE_A) from breast cancer cells were generated separately at four centers (LLU, NCI, FDA and TBU) on four platforms (10x, Fluidigm C1, Fluidigm C1_HT and TBU ICELL8). **c**, Batch-effect correction in scenario 3, in which five scRNA-seq datasets (10X_LLU_B, 10X_NCI_B, C1_FDA_HT_B, C1_LLU_B and ICELL8_SE_B) from B lymphocytes were generated separately at the four centers on the same four platforms. **d**, Batch-effect correction in scenario 4, in which four datasets (10X_LLU_Mix10, 10X_NCI_M_Mix5, 10X_NCI_M_Mix5_F and 10X_NCI_M_Mix5_F2) were generated from 5% or 10% breast cancer cells spiked into B lymphocytes and analyzed with the 10x Genomics platform at two centers in four different batches. Each dataset is indicated by a unique color in panels **a**–**d**. Idealized projections of cells for the four different scenarios are presented on the left. *Note that for the BBKNN analysis, only data generated by UMAP are available and shown; all others are t-SNE plots. Silhouette width score quantifying the clusterability for scenario 1 (**e**) or scenario 4 (**f**), corresponding to **a** and **d**, respectively. **g**, kBET acceptance score quantifying mixability, calculated using the cross-platform, cross-center scRNA-seq data acquired either from breast cancer cells only or from B lymphocytes only from each of the four scenarios (**a**–**d**, also labeled as scenarios 1–4).

nature biotechnology

FOCUS | ANALYSIS

https://doi.org/10.1038/s41587-021-00994-5

Check for updates

Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing

Wenming Xiao[1,44] ✉, Luyao Ren[2,44], Zhong Chen[3], Li Tai Fang[4], Yongmei Zhao[5], Justin Lack[5], Meijian Guan[6], Bin Zhu[7], Erich Jaeger[8], Liz Kerrigan[9], Thomas M. Blomquist[10], Tiffany Hung[11], Marc Sultan[12], Kenneth Idler[13], Charles Lu[13], Andreas Scherer[14,15], Rebecca Kusko[16], Malcolm Moos[17], Chunlin Xiao[18], Stephen T. Sherry[18], Ogan D. Abaan[8,19], Wanqiu Chen[3], Xin Chen[3], Jessica Nordlund[15,20], Ulrika Liljedahl[15,21], Roberta Maestro[15,21], Maurizio Polano[15,21], Jiri Drabek[15,22], Petr Vojta[15,22], Sulev Kõks[15,23,24], Ene Reimann[15,25], Bindu Swapna Madala[26], Timothy Mercer[26], Chris Miller[13], Howard Jacob[13], Tiffany Truong[8], Ali Moshrefi[8], Aparna Natarajan[8], Ana Granat[8], Gary P. Schroth[8], Rasika Kalamegham[11], Eric Peters[11], Virginie Petitjean[12], Ashley Walton[5], Tsai-Wei Shen[5], Keyur Talsania[5], Cristobal Juan Vera[5], Kurt Langenbach[9], Maryellen de Mars[9], Jennifer A. Hipp[10], James C. Willey[10], Jing Wang[27], Jyoti Shetty[28], Yuliya Kriga[28], Arati Raziuddin[28], Bao Tran[28], Yuanting Zheng[2], Ying Yu[2], Margaret Cam[29], Parthav Jailwala[29], Cu Nguyen[30], Daoud Meerzaman[30], Qingrong Chen[30], Chunhua Yan[30], Ben Ernest[31], Urvashi Mehra[31], Roderick V. Jensen[32], Wendell Jones[33], Jian-Liang Li[34], Brian N. Papas[34], Mehdi Pirooznia[35], Yun-Ching Chen[35], Fayaz Seifuddin[35], Zhipan Li[36], Xuelu Liu[37], Wolfgang Resch[37], Jingya Wang[38], Leihong Wu[39], Gokhan Yavas[39], Corey Miles[39], Baitang Ning[39], Weida Tong[39], Christopher E. Mason[40], Eric Donaldson[41], Samir Lababidi[42], Louis M. Staudt[43], Zivana Tezak[1], Huixiao Hong[39], Charles Wang[3] ✉ and Leming Shi[2] ✉

Clinical applications of precision oncology require accurate tests that can distinguish true cancer-specific mutations from errors introduced at each step of next-generation sequencing (NGS). To date, no bulk sequencing study has addressed the effects of cross-site reproducibility, nor the biological, technical and computational factors that influence variant identification. Here we report a systematic interrogation of somatic mutations in paired tumor–normal cell lines to identify factors affecting detection reproducibility and accuracy at six different centers. Using whole-genome sequencing (WGS) and whole-exome sequencing (WES), we evaluated the reproducibility of different sample types with varying input amount and tumor purity, and multiple library construction protocols, followed by processing with nine bioinformatics pipelines. We found that read coverage and callers affected both WGS and WES reproducibility, but WES performance was influenced by insert fragment size, genomic copy content and the global imbalance score (GIV; G > T/C > A). Finally, taking into account library preparation protocol, tumor content, read coverage and bioinformatics processes concomitantly, we recommend actionable practices to improve the reproducibility and accuracy of NGS experiments for cancer mutation detection.

Xiao W, Ren L, Chen Z, Fang LT, Zhao Y, Lack J, Guan M, Zhu B, Jaeger E, Kerrigan L, Blomquist TM, Hung T, Sultan M, Idler K, Lu C, Scherer A, Kusko R, Moos M, Xiao C, Sherry ST, Abaan OD, Chen W, Chen X, Nordlund J, Liljedahl U, Maestro R, Polano M, Drabek J, Vojta P, Kõks S, Reimann E, Madala BS, Mercer T, Miller C, Jacob H, Truong T, Moshrefi A, Natarajan A, Granat A, Schroth GP, Kalamegham R, Peters E, Petitjean V, Walton A, Shen TW, Talsania K, Vera CJ, Langenbach K, de Mars M, Hipp JA, Willey JC, Wang J, Shetty J, Kriga Y, Raziuddin A, Tran B, Zheng Y, Yu Y, Cam M, Jailwala P, Nguyen C, Meerzaman D, Chen Q, Yan C, Ernest B, Mehra U, Jensen RV, Jones W, Li JL, Papas BN, Pirooznia M, Chen YC, Seifuddin F, Li Z, Liu X, Resch W, Wang J, Wu L, Yavas G, Miles C, Ning B, Tong W, Mason CE, Donaldson E, Lababidi S, Staudt LM, Tezak Z, Hong H, **Wang C (co-corresponding)**, Shi L. Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing. doi: 10.1038/s41587-021-00994-5. Epub 2021 Sep 9. PMID: 34504346; PMCID: PMC8506910.

Nature Biotechnology 2021 Sep;39(9):1141-1150

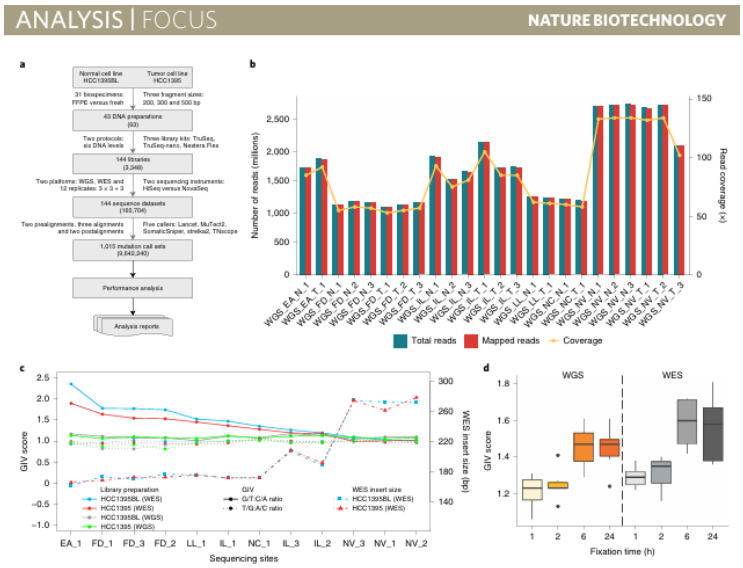


Fig. 1 | Study design and read quality. a, Study design used to capture nonanalytical and analytical factors affecting cancer mutation detection. DNA was extracted from either fresh cells or FFPE-processed cells and fragmented at three intended sizes. Libraries with various levels of DNA input (either from random shotgun or exome capture) were generated with three different library preparation kits and run on WGS and WES in parallel following recommended protocols (Methods). Twelve replicates were performed at six sequencing centers: three centers (FD, IL and NV) prepared WGS and WES libraries in triplicate; three centers (EA, LL and NC) prepared a single WGS and WES library (3 × 3 + 3); and 144 libraries were sequenced on either a HiSeq or NovaSeq instrument. Two prealignments (BFC and Trimmomatic), three alignments (BWA, Bowtie and NovoAlign) and two postalignments (GATK and no-GATK) were evaluated. A total of 1,015 mutation call sets were generated. Numbers in parentheses represent possible combinations at that level. Further details on the experiment design are given in Extended Data Fig. 1. **b**, Read yields (blue), mapping statistics (red) and genome coverage (yellow line) from 12 repeated WGS runs. **c**, GIV of G > T/C > A and T > /A > C mutation pairs in WES and WGS runs. Six centers used a range of time spans (80–300 s) for DNA shearing. As a result, average insert DNA fragment size ranged from 161 to 274 bp. **d**, Distribution of GIV score for FFPE DNA with four different fixation times (1, 2, 6 and 24 h) analyzed with WES or WGS: FFX and FFPE on WES platform and FFG and FFPE on WGS platform. Box-and-whisker plots shows the first and third quartiles as well as median values. The upper and lower whiskers extend from the hinge to the largest or smallest value no further than 1.5 × interquartile range from the hinge. For detailed statistics regarding minima, maxima, center, bounds of box and whiskers, and percentiles related to this figure, please refer to Supplementary Table 4.

nature biotechnology

FOCUS | RESOURCE

https://doi.org/10.1038/s41587-021-00993-6

Check for updates

Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing

Li Tai Fang[1,54], Bin Zhu[2,54], Yongmei Zhao[3,54], Wanqiu Chen[4], Zhaowei Yang[4,5], Liz Kerrigan[6], Kurt Langenbach[6], Maryellen de Mars[6], Charles Lu[7], Kenneth Idler[7], Howard Jacob[7], Yuanting Zheng[8], Luyao Ren[8], Ying Yu[8], Erich Jaeger[9], Gary P. Schroth[9], Ogan D. Abaan[9], Keyur Talsania[3], Justin Lack[3], Tsai-Wei Shen[3], Zhong Chen[4], Seta Stanbouly[4], Bao Tran[10], Jyoti Shetty[10], Yuliya Kriga[10], Daoud Meerzaman[11], Cu Nguyen[11], Virginie Petitjean[12], Marc Sultan[12], Margaret Cam[13], Monika Mehta[10], Tiffany Hung[14], Eric Peters[14], Rasika Kalamegham[14], Sayed Mohammad Ebrahim Sahraeian[1], Marghoob Mohiyuddin[1], Yunfei Guo[1], Lijing Yao[1], Lei Song[2], Hugo Y. K. Lam[1], Jiri Drabek[15,16], Petr Vojta[15,16], Roberta Maestro[16,17], Daniela Gasparotto[16,17], Sulev Kõks[16,18,19], Ene Reimann[16,19], Andreas Scherer[16,20], Jessica Nordlund[16,21], Ulrika Liljedahl[16,21], Roderick V. Jensen[22], Mehdi Pirooznia[23], Zhipan Li[24], Chunlin Xiao[25], Stephen T. Sherry[25], Rebecca Kusko[26], Malcolm Moos[27], Eric Donaldson[28], Zivana Tezak[29], Baitang Ning[30], Weida Tong[30], Jing Li[5], Penelope Duerken-Hughes[31], Claudia Catalanotti[32], Shamoni Maheshwari[32], Joe Shuga[32], Winnie S. Liang[33], Jonathan Keats[33], Jonathan Adkins[33], Erica Tassone[33], Victoria Zismann[33], Timothy McDaniel[33], Jeffrey Trent[33], Jonathan Foox[34], Daniel Butler[34], Christopher E. Mason[34], Huixiao Hong[30] ✉, Leming Shi[8] ✉, Charles Wang[4,31] ✉, Wenming Xiao[29] ✉ and The Somatic Mutation Working Group of Sequencing Quality Control Phase II Consortium*

The lack of samples for generating standardized DNA datasets for setting up a sequencing pipeline or benchmarking the performance of different algorithms limits the implementation and uptake of cancer genomics. Here, we describe reference call sets obtained from paired tumor–normal genomic DNA (gDNA) samples derived from a breast cancer cell line—which is highly heterogeneous, with an aneuploid genome, and enriched in somatic alterations—and a matched lymphoblastoid cell line. We partially validated both somatic mutations and germline variants in these call sets via whole-exome sequencing (WES) with different sequencing platforms and targeted sequencing with >2,000-fold coverage, spanning 82% of genomic regions with high confidence. Although the gDNA reference samples are not representative of primary cancer cells from a clinical sample, when setting up a sequencing pipeline, they not only minimize potential biases from technologies, assays and informatics but also provide a unique resource for benchmarking 'tumor-only' or 'matched tumor–normal' analyses.

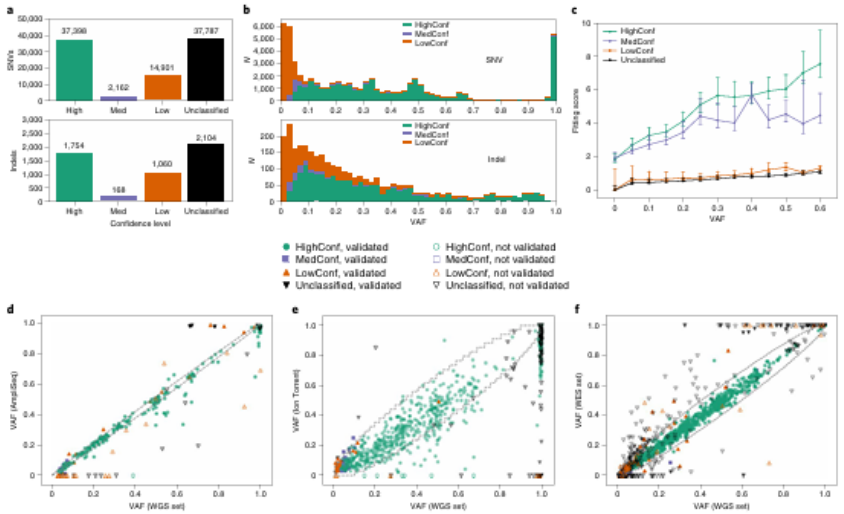


Fig. 2 | Definition and validation of the somatic mutation reference call set. a, Breakdown of the somatic variant calls within the consensus callable regions based on the four labels HighConf, MedConf, LowConf and Unclassified. Variant calls labeled HighConf and MedConf are grouped into the reference call set; genomic positions with LowConf and Unclassified calls are removed from the high-confidence regions. **b**, Histogram of VAFs of the somatic variant calls. **c**, Average tumor purity fitting scores with 95% confidence intervals for the VAF of each SNV across the four different confidence levels versus the observed VAF in the tumor–normal titration series. The formula for fitting scores is described in equation 1 (see Methods for details). **d**, Scatter plot of VAFs observed in 21 WGS datasets versus an AmpliSeq targeted sequencing dataset. Solid shapes represent variants that were validated. Open shapes represent variants that were not validated. Sticks represent uninterpretable validation data. The diagonal dashed lines represent the 95% binomial confidence interval of the observed VAF given the actual VAF calculated based on 2,000× depth for AmpliSeq. The figure shows a very high correlation between VAFs estimated from the WGS data and AmpliSeq data for HighConf calls (Pearson's $R = 0.982$). Many Unclassified data points lie at the bottom, implying that these calls were not real mutations, despite the large number of apparent variant-supporting reads in the all-inclusive set data; *x* axis, VAFs calculated from the all-inclusive set; *y* axis, VAFs calculated from the AmpliSeq data. **e**, Scatter plot of VAFs observed in WGS datasets versus Ion Torrent WES. The 95% binomial confidence intervals were calculated based on 34× depth for Ion Torrent. Pearson's $R = 0.930$ for HighConf calls. **f**, Scatter plot of VAFs observed in WGS datasets versus 12 repeats of WES on the HiSeq platform; *y* axis, median VAFs calculated based on 12 HiSeq WES replicates. The 95% binomial confidence intervals were calculated based on 150× depth for HiSeq WES. Pearson's $R = 0.997$ for HighConf calls. In **d**–**f**, the colors indicate the confidence level of the variant calls, whereas the shapes indicate their validation status.

Fang LT, Zhu B, Zhao Y, Chen W, Yang Z, Kerrigan L, Langenbach K, de Mars M, Lu C, Idler K, Jacob H, Zheng Y, Ren L, Yu Y, Jaeger E, Schroth GP, Abaan OD, Talsania K, Lack J, Shen TW, Chen Z, Stanbouly S, Tran B, Shetty J, Kriga Y, Meerzaman D, Nguyen C, Petitjean V, Sultan M, Cam M, Mehta M, Hung T, Peters E, Kalamegham R, Sahraeian SME, Mohiyuddin M, Guo Y, Yao L, Song L, Lam HYK, Drabek J, Vojta P, Maestro R, Gasparotto D, Kõks S, Reimann E, Scherer A, Nordlund J, Liljedahl U, Jensen RV, Pirooznia M, Li Z, Xiao C, Sherry ST, Kusko R, Moos M, Donaldson E, Tezak Z, Ning B, Tong W, Li J, Duerken-Hughes P, Catalanotti C, Maheshwari S, Shuga J, Liang WS, Keats J, Adkins J, Tassone E, Zismann V, McDaniel T, Trent J, Foox J, Butler D, Mason CE, Hong H, Shi L, **Wang C** (**co-corresponding**), Xiao W; Somatic Mutation Working Group of Sequencing Quality Control Phase II Consortium. Establishing community reference samples, data and call sets for benchmarking cancer mutation detection using whole-genome sequencing. doi: 10.1038/s41587-021-00993-6. Epub 2021 Sep 9. PMID: 34504347; PMCID: PMC8532138.

Nature Biotechnology 2021 Sep;39(9):1151-1160

Precision Clinical Medicine, 4(4), 2021, 231–245

https://doi.org/10.1093/pcmedi/pbab023
Advance Access Publication Date: 28 August 2021
Research Article

RESEARCH ARTICLE

Inhibitory activity of medicinal mushroom *Ganoderma lucidum* on colorectal cancer by attenuating inflammation

Mandy M. Liu[1,§], Tiantian Liu[2,§], Steven Yeung[1], Zhijun Wang[3], Bradley Andresen[1], Cyrus Parsa[4,5], Robert Orlando[4,5], Bingsen Zhou[6], Wei Wu[6], Xia Li[6], Yilong Zhang[6], Charles Wang[2,*] and Ying Huang[1,*]

[1]Department of Pharmaceutical Sciences, College of Pharmacy, Western University of Health Sciences, Pomona, CA 91766, USA
[2]Center for Genomics & Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[3]Department of Pharmaceutical Sciences, College of Pharmacy, Marshall B. Ketchum University, Fullerton, CA 92831, USA
[4]College of Osteopathic Medicine of the Pacific, Western University of Health Sciences, Pomona, CA 91766, USA
[5]Department of Pathology, Beverly Hospital, Montebello, California, CA 90640, USA
[6]Beijing Tong Ren Tang Chinese Medicine Co., Ltd., New Territories, Hong Kong 999077, China

*Correspondence: Charles Wang, chwang@llu.edu; Ying Huang, yhuang@westernu.edu
Charles Wang, http://orcid.org/0000-0001-8861-2121
§Mandy M. Liu and Tiantian Liu contributed equally to this work.

Abstract

The medicinal mushroom *Ganoderma lucidum* (GL, Reishi or Lingzhi) exhibits an inhibitory effect on cancers. However, the underlying mechanism of the antitumor activity of GL is not fully understood. In this study, we characterized the gene networks regulated by a commercial product of GL containing a mixture of spores and fruiting bodies namely "GLSF", in colorectal carcinoma. We found that *in vitro* co-administration of GLSF extract at non-toxic concentrations significantly potentiated growth inhibition and apoptosis induced by paclitaxel in CT26 and HCT-15 cells. GLSF inhibited NF-κB promoter activity in HEK-293 cells but did not affect the function of P-glycoprotein in K562/DOX cells. Furthermore, we found that when mice were fed a modified diet containing GLSF for 1 month prior to the CT26 tumor cell inoculation, GLSF alone or combined with Nab-paclitaxel markedly suppressed tumor growth and induced apoptosis. RNA-seq analysis of tumor tissues derived from GLSF-treated mice identified 53 differentially expressed genes compared to normal tissues. Many of the GLSF-down-regulated genes were involved in NF-κB-regulated inflammation pathways, such as IL-1β, IL-11 and Cox-2.

Received: 2 June 2021; **Revised:** 16 August 2021; **Accepted:** 23 August 2021

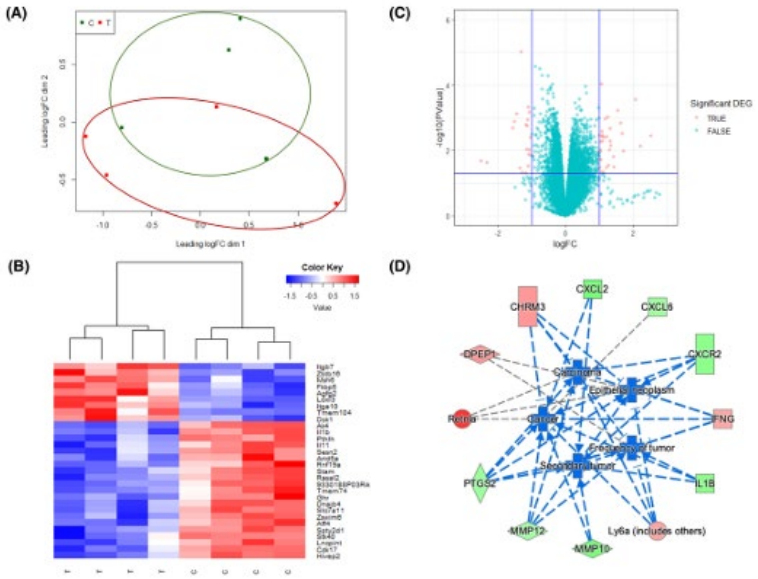


Figure 5. RNA-seq analysis of CT26 syngeneic tumors in mice treated with control diet or GLSF-modified diet. (a) Multi-dimensional scaling plot of detected genes in GLSF treatment (T, red) and control group (C, green). The distances correspond to leading log2 fold-changes between each pair of samples. (b) Volcano plots of differentially expressed genes between treatment and control groups. The red dots indicate up- and down-regulated DEGs with $P < 0.05$ and absolute log2 FC > 1. (c) Heat map of top 30 significantly differentially expressed genes between treatment (T) and control groups (C). (d) Genes that are up- and down-regulated in treatment group (compared to control) are displayed within red or green nodes, respectively. The predicted inhibited biology effects are presented in blue nodes. Blue (predicted to be inhibited) or gray (undetermined direction) dash lines represent relationships with causal consistency.

Liu MM, Liu T, Yeung S, Wang Z, Andresen B, Parsa C, Orlando R, Zhou B, Wu W, Li X, Zhang Y, **Wang C** (**co-corresponding**), Huang Y. Inhibitory activity of medicinal mushroom *Ganoderma lucidum* on colorectal cancer by attenuating inflammation. doi: 10.1093/pcmedi/pbab023. PMID: 35692861; PMCID: PMC8982591.

Precision Clinical Medicine 2021 Aug 28;4(4):231-245

www.nature.com/scientificreports

scientific reports

OPEN

Novel genomic targets of valosin-containing protein in protecting pathological cardiac hypertrophy

Ning Zhou[1,2,6], Xin Chen[3,6], Jing Xi[1,6], Ben Ma[1,5], Christiana Leimena[1], Shaunrick Stoll[1], Gangjian Qin[4], Charles Wang[3]✉ & Hongyu Qiu[1,5]✉

Pressure overload-induced cardiac hypertrophy, such as that caused by hypertension, is a key risk factor for heart failure. However, the underlying molecular mechanisms remain largely unknown. We previously reported that the valosin-containing protein (VCP), an ATPase-associated protein newly identified in the heart, acts as a significant mediator of cardiac protection against pressure overload-induced pathological cardiac hypertrophy. Still, the underlying molecular basis for the protection is unclear. This study used a cardiac-specific VCP transgenic mouse model to understand the transcriptomic alterations induced by VCP under the cardiac stress caused by pressure overload. Using RNA sequencing and comprehensive bioinformatic analysis, we found that overexpression of the VCP in the heart was able to normalize the pressure overload-stimulated hypertrophic signals by activating G protein-coupled receptors, particularly, the olfactory receptor family, and inhibiting the transcription factor controlling cell proliferation and differentiation. Moreover, VCP overexpression restored pro-survival signaling through regulating alternative splicing alterations of mitochondrial genes. Together, our study revealed a novel molecular regulation mediated by VCP under pressure overload that may bring new insight into the mechanisms involved in protecting against hypertensive heart failure.

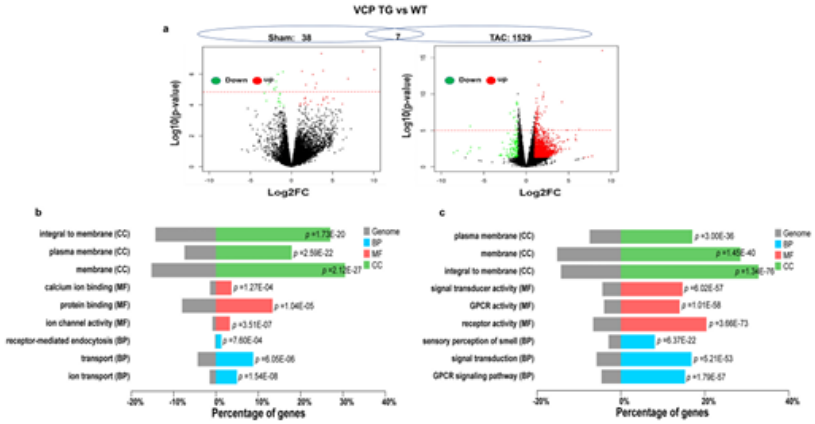


Figure 1. VCP-mediated transcriptomic alterations are different between the sham and 2 weeks (2W) TAC conditions. (**a**) The volcano plot of DEGs between VCP TG and WT mice at the sham and 2W TAC conditions. Red and green dots represent up- and down-regulated in the VCP TG group, respectively. The red dash line represents the threshold of FDR. The dot above the line are genes with FDR < 0.05. n = 3–4/group. (**b**), (**c**) GO functional analysis of DEGs induced by VCP between the sham and TAC conditions. GO functions of DEGs based on the FC > Log2 and $p < 0.05$ by comparing VCP TG with WT mice at the condition of sham (**b**) and 2W TAC (**c**), in terms of the cellular components (CC, green), molecular function (MF, red) and biological process (BP, Blue).

Zhou N, Chen X, Xi J, Ma B, Leimena C, Stoll S, Qin G, **Wang C (co-corresponding)**, Qiu H. Novel genomic targets of valosin-containing protein in protecting pathological cardiac hypertrophy. doi: 10.1038/s41598-020-75128-z. PMID: 33093614; PMCID: PMC7582185.

Scientific Reports 2020 Oct 22;10(1):18098

OXFORD

Precision Clinical Medicine, 3(4), 2020, 245–259

doi: 10.1093/pcmedi/pbaa037
Advance Access Publication Date: 23 October 2020
Research Article

RESEARCH ARTICLE

Vegetarian diets, circulating miRNA expression and healthspan in subjects living in the Blue Zone

Tiantian Liu[1], Nicole M. Gatto[2,§], Zhong Chen[1,§], Hongyu Qiu[3], Grace Lee[4], Penelope Duerksen-Hughes[5], Gary Fraser[6] and Charles Wang[1,5,*]

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[2]School of Community and Global Health, Claremont Graduate University, Claremont, CA 91711, USA
[3]Center of Molecular and Translational Medicine, Institution of Biomedical Science, Georgia State University, Atlanta, GA 30303, USA
[4]Department of Psychology, School of Behavioral Health, Loma Linda University, Loma Linda, CA 92350, USA
[5]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA
[6]School of Public Health, Loma Linda University, Loma Linda, CA 92350, USA

*Correspondence: Charles Wang, oxwang@gmail.com or chwang@llu.edu
Charles Wang, http://orcid.org/0000-0001-8861-2121
§These authors contributed equally to this work.

Abstract

A long-term vegetarian diet plays a role in the longevity and maintenance of the healthspan, but the underlying mechanisms for these observations are largely unknown. Particularly, it is not known whether a long-term vegetarian dietary pattern may affect the circulating miRNA expression in such a way as to modulate the healthspan. The Adventist Health Study-2 (AHS-2) cohort includes a large number of older adults who primarily follow vegetarian dietary patterns and reside in Loma Linda, California, one of five "Blue Zones" in the world in which a higher proportion of the population enjoys a longer than average lifespan. We performed miRNA-seq in 96 subjects selected from the AHS-2 cohort with different dietary patterns. We identified several differentially expressed miRNAs between vegetarians and non-vegetarians, which are involved in immune response and cytokine signaling, cell growth and proliferation as well as age-related diseases such as cardiovascular diseases and neurodegenerative diseases. Overall, our study showed that a vegetarian diet modulates aging-associated circulating miRNAs in a sex-dependent manner of differential expression for certain miRNAs, which may be related in a beneficial manner to the healthspan. Further investigation is needed to validate these miRNAs as potential biomarkers for diet-modulated longevity in humans.

Key words: circulating miRNA; aging; miRNA sequencing; vegetarian diet; Blue Zone

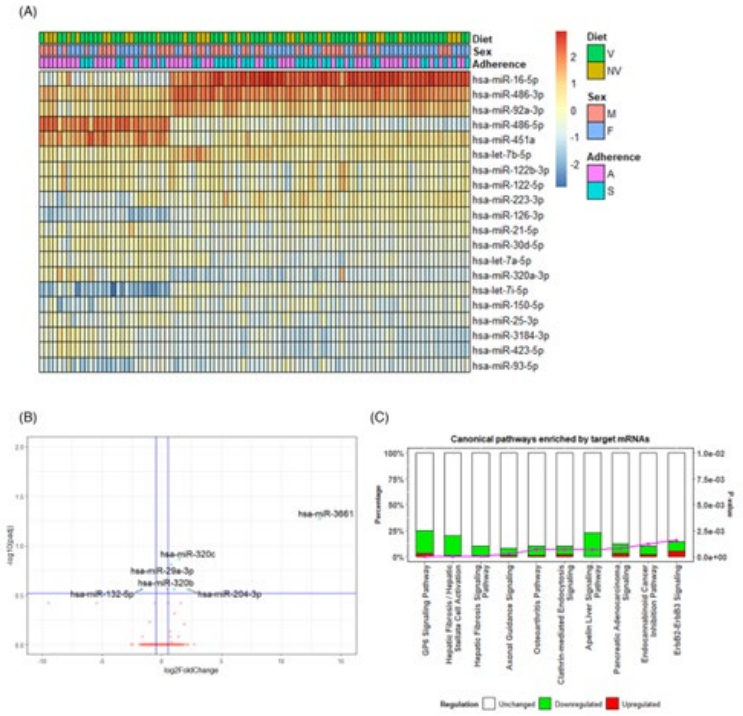


Figure 2. Most abundant miRNAs across all samples and differentially expressed miRNAs between vegetarians and non-vegetarians. (**A**) Heat map showing the top 20 most abundant miRNAs across all samples; (**B**) Volcano plot showing the differentially expressed miRNAs between V and NV groups. The Y-axis represents the -log10(padj) and the X-axis displays the log2FC value. A positive x-value represents an upregulation and a negative x-value represents a downregulation compared to NV. -log10(padj) ≥ 0.52 (padj ≤ 0.3) and |log2FC| ≥ 0.5 were marked as the significant threshold. Each dot represents one differentially expressed miRNA, with those above the significance threshold highlighted in blue. Significant differentially expressed miRNAs are annotated in the plot. V: Vegetarians, NV: Non-vegetarians. (**C**) Bar graphs showing the top 10 enriched canonical pathways by IPA (V vs. NV). Y-axis[1] (left) represents the percentage of regulated genes to all genes in the pathway. Y-axis[2] (right) represents the P-value of the enrichment. Downregulated genes are colored in green, upregulated genes are colored in red and unchanged genes are colored in white. V: vegetarians; NV: non-vegetarians, M: males; F: females; A: adherers; S: switchers.

Liu T, Gatto NM, Chen Z, Qiu H, Lee G, Duerksen-Hughes P, Fraser G, **Wang C (corresponding)**. Vegetarian diets, circulating miRNA expression and healthspan in subjects living in the Blue Zone. doi: 10.1093/pcmedi/pbaa037. Epub 2020 Oct 23. PMID: 33391847; PMCID: PMC7757436.

Precision Clinical Medicine 2020 Dec;3(4):245-259

ESC European Society of Cardiology
Cardiovascular Research (2019) **115**, 1306–1319
doi:10.1093/cvr/cvy277

Foetal hypoxia impacts methylome and transcriptome in developmental programming of heart disease

Lei Huang[1†], Xin Chen[2†], Chiranjib Dasgupta[1], Wanqiu Chen[2], Rui Song[1], Charles Wang[2]*, and Lubo Zhang[1]*

[1]Department of Basic Sciences, Lawrence D. Longo, MD Center for Perinatal Biology, Loma Linda University School of Medicine, 11234 Anderson Street, Loma Linda, CA 92350, USA and [2]Department of Basic Sciences, Center for Genomics, Loma Linda University School of Medicine, 11234 Anderson Street, Loma Linda, CA 92350, USA

Received 8 July 2018; revised 6 September 2018; editorial decision 31 October 2018; accepted 1 November 2018; online publish-ahead-of-print 5 November 2018

Time for primary review: 10 days

Aims Antenatal hypoxia negatively impacts foetal heart development, and increases the risk of heart disease later in life. The molecular mechanisms remain largely elusive. Here, we conducted a genome-wide analysis to study the impact of antenatal hypoxia on DNA methylome and transcriptome profiling in foetal and adult offspring hearts.

Methods and results Pregnant rats were treated with normoxia or hypoxia (10.5% O_2) from Day 15 to Day 21 of gestation. Hearts were isolated from near-term foetuses and 5-month-old male and female offsprings, and DNA methylome and RNA-seq were performed. Methylome data shows a sharp dip in CpG methylation centred at the transcription start site (TSS). CpG islands (CGIs) and CpG island shores (CGSs) within 10 kb upstream of the TSS are hypomethylated, compared with CGIs and CGSs within gene bodies. Combining transcriptome, data indicate an inverse relation between gene expression and CpG methylation around the TSS. Of interest, antenatal hypoxia induces opposite changes in methylation patterns in foetal and adult hearts, with hypermethylation in the foetus and hypomethylation in the adult. Also, there is significant sex dimorphism of changes in gene expression patterns in the adult offspring heart. Notably, pathway analysis indicates that enrichment of inflammation-related pathways are significantly greater in the adult male heart than those in the female heart.

Conclusion Our study provides an initial framework and new insights into foetal hypoxia-mediated epigenetic programming of pro-inflammatory phenotype in the heart development, linking antenatal stress, and developmental programming of heart vulnerability to disease later in life.

Keywords Antenatal hypoxia • Heart • Methylome • Transcriptome • Inflammation

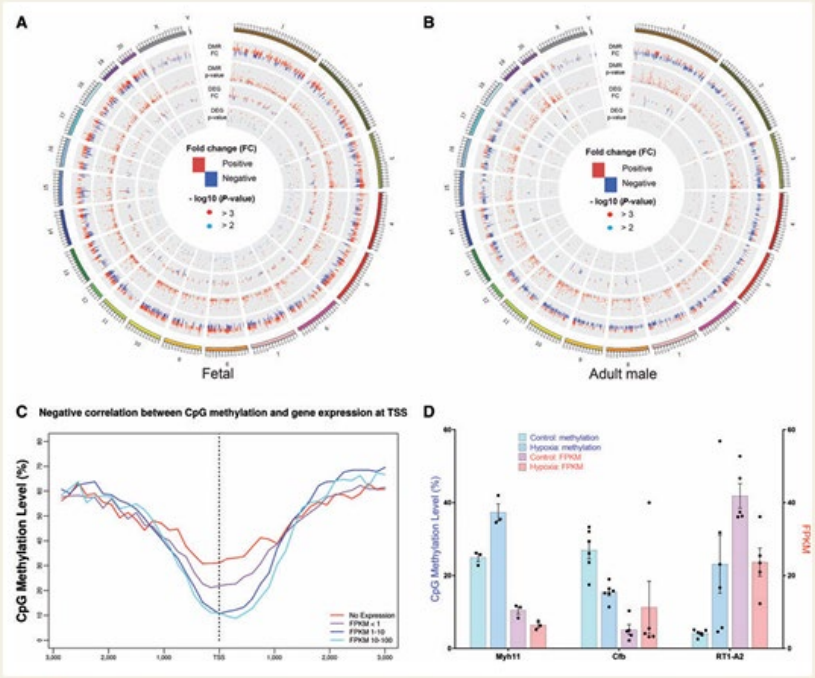


Figure 3 Integration of DMRs and DEGs. (*A*, *B*) Circos plot of DMRs and DEGs in foetal and adult male rats. Positive fold-changes in DMRs represent hypermethylation in hypoxia group while negative fold-changes in DMRs represent hypermethylation in control group. Positive fold-changes in DEGs represent up-regulation in hypoxia group while negative fold-changes in DEGs represent up-regulation in control group. Red dots represent *P*-value <0.001 while blue dots represent 0.001 ≤ *P*-value <0.01 ($n = 3$–6). (*C*) Negative correlation between CpG methylation and gene expression at TSS. (*D*) Three example genes (*Myh11*, *Cfb*, and *RT1-A2*) with DMRs located in promoter regions show a negative correlation between gene expression (FPKM) and methylation rate (beta value) of CpG regions under control and hypoxia environment. Data are expressed as mean ± SEM. $n = 3$–6.

Huang L, Chen X, Dasgupta C, Chen W, Song R, **Wang C (co-corresponding)**, Zhang L. Foetal hypoxia impacts methylome and transcriptome in developmental programming of heart disease. doi: 10.1093/cvr/cvy277. PMID: 30395198; PMCID: PMC6587923.

Cardiovascular Research 2019 Jul 1;115(8):1306-1319.

Precision Clinical Medicine, 2(4), 2019, 221–234

doi: 10.1093/pcmedi/pbz028
Advance Access Publication Date: 8 December 2019
Research Article

RESEARCH ARTICLE

DNA methylome study of human cerebellar tissues identified genes and pathways possibly involved in essential tremor

Jennifer L. Paul[1], Khashayar Dashtipour[2], Zhong Chen[1] and Charles Wang[1,3*]

[1]Center for Genomics, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA; [2]Division of Movement Disorders, Department of Neurology, Loma Linda University Medical Center, Loma Linda, CA 92350, USA; [3]Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, CA 92350, USA

*Correspondence: Charles Wang, chwang@llu.edu

Abstract

Background: Essential tremor (ET) is a neurological syndrome of unknown origin with poorly understood etiology and pathogenesis. It is suggested that the cerebellum and its tracts may be involved in the pathophysiology of ET. DNA methylome interrogation of cerebellar tissue may help shine some light on the understanding of the mechanism of the development of ET. Our study used postmortem human cerebellum tissue samples collected from 12 ET patients and 11 matched non-ET controls for DNA methylome study to identify differentially methylated genes in ET.

Results: Using Nugen's Ovation reduced representation bisulfite sequencing (RRBS), we identified 753 genes encompassing 938 CpG sites with significant differences in DNA methylation between the ET and the control group. Identified genes were further analyzed with Ingenuity Pathway Analysis (IPA) by which we identified certain significant pathways, upstream regulators, diseases and functions, and networks associated with ET.

Conclusions: Our study provides evidence that there are significant differences in DNA methylation patterns between the ET and control samples, suggesting that the methylation alteration of certain genes in the cerebellum may be associated with ET pathogenesis. The identified genes allude to the GABAergic hypothesis which supports the notation that ET is a neurodegenerative disease, particularly involving the cerebellum.

Key words: DNA methylome; reduced representation bisulfite sequencing; essential tremor; cerebellum; epigenetics; CpGs

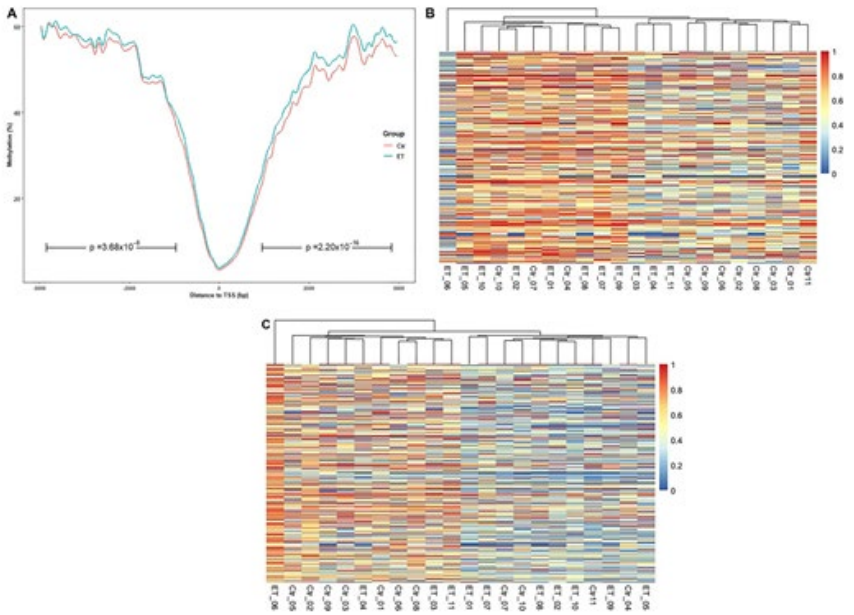


Figure 3. Visualization of methylation difference between ET and control. (**A**) Methylation level of CpGs flanking to TSS. All CpGs annotated within 5 kb of TSS were included. The region from ≥ 1 kb upstream and region from ≤ -1 kb downstream are significantly different. N = 12 in ET group and N = 11 in control group. Heat map of hypermethylated CpG (**B**) and hypomethylated CpG. (**C**) loci. Only CpGs that are annotated to known genes were selected for the heat map plot. Due to space limitation, the CpG information is not shown on the plot, but listed in Supplementary File 3. Hierarchical cluster analysis on sample dissimilarity was calculated using the "complete" method.

Paul JL, Dashtipour K, Chen Z, **Wang C (corresponding)**. DNA methylome study of human cerebellar tissues identified genes and pathways possibly involved in essential tremor. doi: 10.1093/pcmedi/pbz028. Epub 2019 Dec 8. PMID: 31886034; PMCID: PMC6927097.

Precision Clinical Medicine 2019 Dec;2(4):221-234

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Epigenomic Reprogramming of Adult Cardiomyocyte-Derived Cardiac Progenitor Cells

Received: 13 March 2015
Accepted: 14 October 2015
Published: 14 December 2015

Yiqiang Zhang[1,2], Jiang F. Zhong[3], Hongyu Qiu[4], W. Robb MacLellan[1], Eduardo Marbán[2] & Charles Wang[5]

It has been believed that mammalian adult cardiomyocytes (ACMs) are terminally-differentiated and are unable to proliferate. Recently, using a bi-transgenic ACM fate mapping mouse model and an *in vitro* culture system, we demonstrated that adult mouse cardiomyocytes were able to dedifferentiate into cardiac progenitor-like cells (CPCs). However, little is known about the molecular basis of their intrinsic cellular plasticity. Here we integrate single-cell transcriptome and whole-genome DNA methylation analyses to unravel the molecular mechanisms underlying the dedifferentiation and cell cycle reentry of mouse ACMs. Compared to parental cardiomyocytes, dedifferentiated mouse cardiomyocyte-derived CPCs (mCPCs) display epigenomic reprogramming with many differentially-methylated regions, both hypermethylated and hypomethylated, across the entire genome. Correlated well with the methylome, our transcriptomic data showed that the genes encoding cardiac structure and function proteins are remarkably down-regulated in mCPCs, while those for cell cycle, proliferation, and stemness are significantly up-regulated. In addition, implantation of mCPCs into infarcted mouse myocardium improves cardiac function with augmented left ventricular ejection fraction. Our study demonstrates that the cellular plasticity of mammalian cardiomyocytes is the result of a well-orchestrated epigenomic reprogramming and a subsequent global transcriptomic alteration.

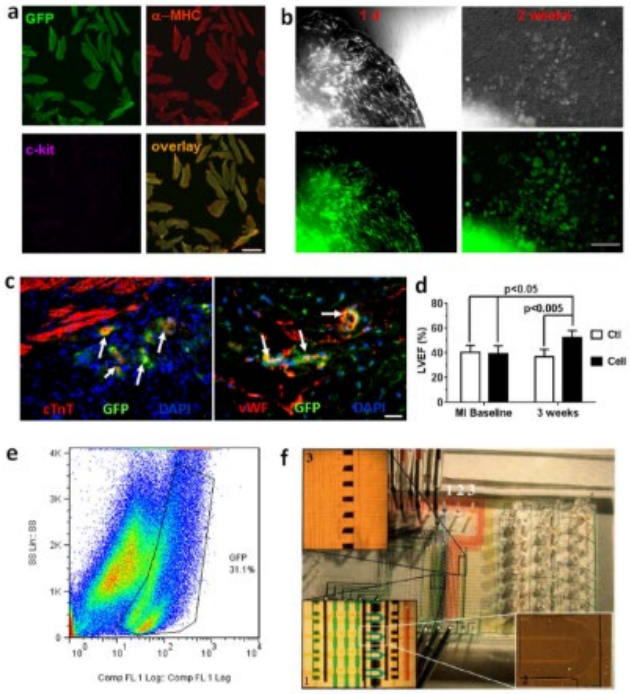


Figure 1. Cardiomyocyte dedifferentiation, re-differentiation, and isolation of single cells for whole-transcriptome analysis. (**a**) Immunocytochemistry showing cardiac-specific labeling of GFP (green) in mature myocytes expressing α-MHC (red) but not c-kit (magenta). Adult cardiomyocytes were isolated from bi-transgenic αMHC-MerCreMer Z/EG mouse hearts after tamoxifen-induced gene recombination (see Methods for detail). Scale bar, 200 μm. (**b**) Phase contrast images (upper panels) and fluorescent images (lower panels, Green- GFP) of transgenic cardiac tissue and derived cells at the beginning of culture (1d) and at 2 weeks of culture. Scale bar, 200 μm. (**c**) Confocal microscope images showing the re-differentiation of GFP-mCPCs (green) into cardiomyocytes expressing cardiac troponin (cTnT, red; white arrows, left panel) and endothelial cells expressing von Willebrand factor (VWF, red; white arrows, right panel) 3 weeks after implantation into the infarcted myocardium of wild-type background mice. Nuclei were stained with DAPI (blue). Scale bar, 100 μm. (**d**) Left ventricular ejection fraction (LVEF) measured at baseline after myocardial infarction (MI) surgery and 3 week post-MI using echocardiography. Vehicle media (Ctl) was used as a control for GFP-mCPC cell transplantation (Cell). N = 5 mice for Ctl group, and 6 mice for Cell group. (**e**) Flow cytometry analysis of cardiac explant cultures at 2 weeks showing about 30% cells are GFP positive in the whole population. (**f**) Custom phase-switch microfluidic chip for the isolation of single mCPCs. Individual live single-cell can be encapsulated into as little as 500-pl droplet for downstream analysis.

Zhang Y, Zhong JF, Qiu H, MacLellan WR, Marbán E, **Wang C (corresponding)**. Epigenomic Reprogramming of Adult Cardiomyocyte-Derived Cardiac Progenitor Cells. doi: 10.1038/srep17686. Erratum in: Sci Rep. 2017 Oct 20;7:46907. PMID: 26657817; PMCID: PMC4677315.

Scientific Reports 2015 Dec 14;5:17686

nature COMMUNICATIONS

ARTICLE

Received 7 Jun 2013 | Accepted 10 Jan 2014 | Published 10 Feb 2014

DOI: 10.1038/ncomms4230 OPEN

A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages

Ying Yu[1,*], James C. Fuscoe[2,*], Chen Zhao[1], Chao Guo[3], Meiwen Jia[1], Tao Qing[1], Desmond I. Bannon[4], Lee Lancashire[5], Wenjun Bao[6], Tingting Du[1], Heng Luo[1], Zhenqiang Su[2], Wendell D. Jones[7], Carrie L. Moland[2], William S. Branham[2], Feng Qian[2], Baitang Ning[2], Yan Li[2], Huixiao Hong[2], Lei Guo[2], Nan Mei[2], Tieliu Shi[8], Kevin Y. Wang[9], Russell D. Wolfinger[6], Yuri Nikolsky[5], Stephen J. Walker[10], Penelope Duerksen-Hughes[11], Christopher E. Mason[12], Weida Tong[2], Jean Thierry-Mieg[13], Danielle Thierry-Mieg[13], Leming Shi[1,2,14] & Charles Wang[15]

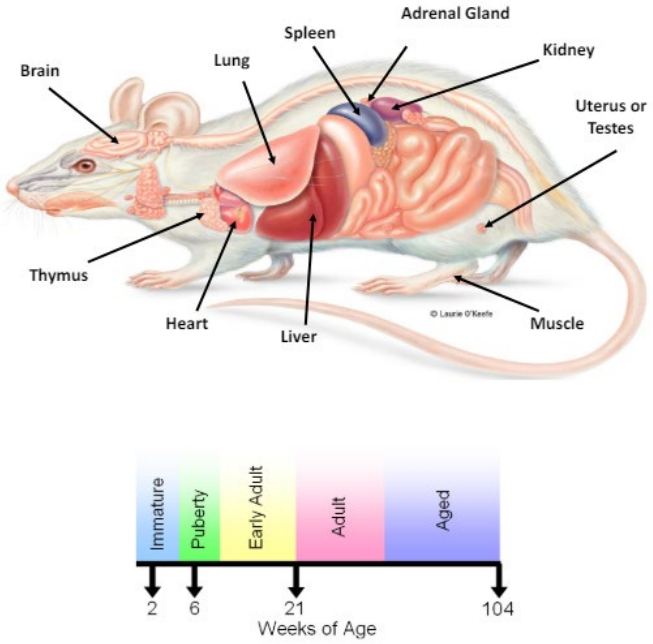
The rat has been used extensively as a model for evaluating chemical toxicities and for understanding drug mechanisms. However, its transcriptome across multiple organs, or developmental stages, has not yet been reported. Here we show, as part of the SEQC consortium efforts, a comprehensive rat transcriptomic BodyMap created by performing RNA-Seq on 320 samples from 11 organs of both sexes of juvenile, adolescent, adult and aged Fischer 344 rats. We catalogue the expression profiles of 40,064 genes, 65,167 transcripts, 31,909 alternatively spliced transcript variants and 2,367 non-coding genes/non-coding RNAs (ncRNAs) annotated in AceView. We find that organ-enriched, differentially expressed genes reflect the known organ-specific biological activities. A large number of transcripts show organ-specific, age-dependent or sex-specific differential expression patterns. We create a web-based, open-access rat BodyMap database of expression profiles with crosslinks to other widely used databases, anticipating that it will serve as a primary resource for biomedical research using the rat model.

Yu Y, Fuscoe JC, Zhao C, Guo C, Jia M, Qing T, Bannon DI, Lancashire L, Bao W, Du T, Luo H, Su Z, Jones WD, Moland CL, Branham WS, Qian F, Ning B, Li Y, Hong H, Guo L, Mei N, Shi T, Wang KY, Wolfinger RD, Nikolsky Y, Walker SJ, Duerksen-Hughes P, Mason CE, Tong W, Thierry-Mieg J, Thierry-Mieg D, Shi L, **Wang C (corresponding)**. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. doi: 10.1038/ncomms4230. PMID: 24510058; PMCID: PMC3926002.

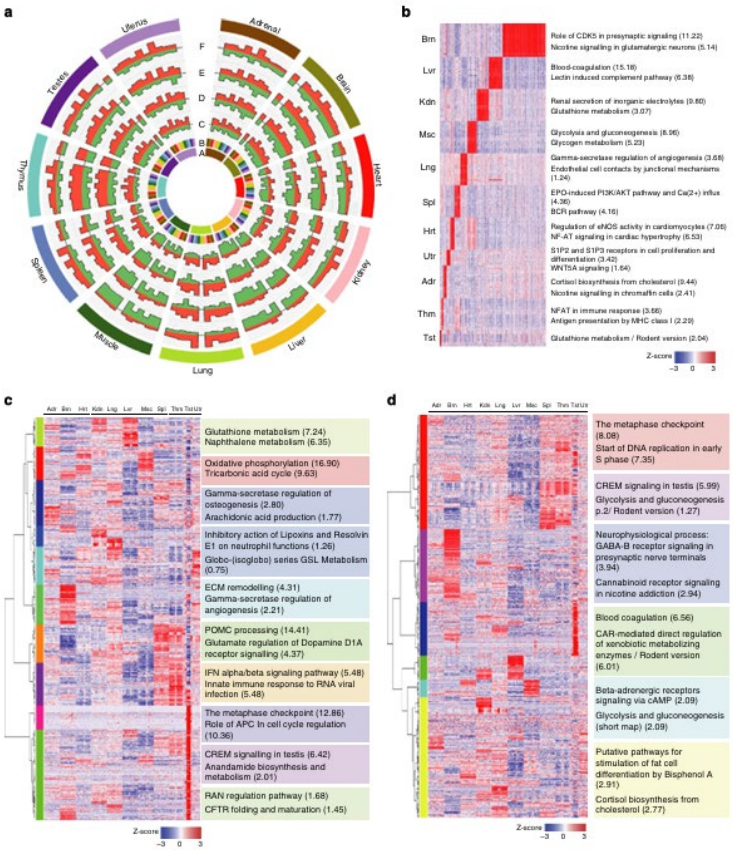
Nature Communications 2014;5:3230

A rat RNA-Seq transcriptomic Bodymap across 11 organs and 4 developmental stages. Yu Y *et al.*

Supplementary Figures and Tables



Supplementary Figure S1. Rat transcriptomic BodyMap study design. Diagram of organs harvested from Fischer 344 rats and the selected four developmental stages at which organs were harvested. Total number of RNA samples: 10 organs per rat x 32 rats (2 sexes x 4 developmental stages x 4 replicates) = 320 for RNA-Seq.



ARTICLES

nature biotechnology

The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance

Charles Wang[1,27], Binsheng Gong[2,27], Pierre R Bushel[3,4,27], Jean Thierry-Mieg[5], Danielle Thierry-Mieg[5], Joshua Xu[2], Hong Fang[6], Huixiao Hong[2], Jie Shen[2], Zhenqiang Su[2], Joe Meehan[2], Xiaojin Li[7], Lu Yang[7], Haiqing Li[7], Paweł P Łabaj[8], David P Kreil[8,9], Dalila Megherbi[10], Stan Gaj[11], Florian Caiment[11], Joost van Delft[11], Jos Kleinjans[11], Andreas Scherer[12], Viswanath Devanarayan[13], Jian Wang[14], Yong Yang[14], Hui-Rong Qian[14], Lee J Lancashire[15], Marina Bessarabova[15], Yuri Nikolsky[16], Cesare Furlanello[17], Marco Chierici[17], Davide Albanese[17,18], Giuseppe Jurman[17], Samantha Riccadonna[17,18], Michele Filosi[17], Roberto Visintainer[17], Ke K Zhang[19], Jianying Li[3,20], Jui-Hua Hsieh[21], Daniel L Svoboda[22], James C Fuscoe[23], Youping Deng[24], Leming Shi[2,25], Richard S Paules[26], Scott S Auerbach[21] & Weida Tong[2]

The concordance of RNA-sequencing (RNA-seq) with microarrays for genome-wide analysis of differential gene expression has not been rigorously assessed using a range of chemical treatment conditions. Here we use a comprehensive study design to generate Illumina RNA-seq and Affymetrix microarray data from the same liver samples of rats exposed in triplicate to varying degrees of perturbation by 27 chemicals representing multiple modes of action (MOAs). The cross-platform concordance in terms of differentially expressed genes (DEGs) or enriched pathways is linearly correlated with treatment effect size ($R^2 \approx 0.8$). Furthermore, the concordance is also affected by transcript abundance and biological complexity of the MOA. RNA-seq outperforms microarray (93% versus 75%) in DEG verification as assessed by quantitative PCR, with the gain mainly due to its improved accuracy for low-abundance transcripts. Nonetheless, classifiers to predict MOAs perform similarly when developed using data from either platform. Therefore, the endpoint studied and its biological complexity, transcript abundance and the genomic application are important factors in transcriptomic research and for clinical and regulatory decision making.

Wang C, Gong B, Bushel PR, Thierry-Mieg J, Thierry-Mieg D, Xu J, Fang H, Hong H, Shen J, Su Z, Meehan J, Li X, Yang L, Li H, Łabaj PP, Kreil DP, Megherbi D, Gaj S, Caiment F, van Delft J, Kleinjans J, Scherer A, Devanarayan V, Wang J, Yang Y, Qian HR, Lancashire LJ, Bessarabova M, Nikolsky Y, Furlanello C, Chierici M, Albanese D, Jurman G, Riccadonna S, Filosi M, Visintainer R, Zhang KK, Li J, Hsieh JH, Svoboda DL, Fuscoe JC, Deng Y, Shi L, Paules RS, Auerbach SS, Tong W. The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. doi: 10.1038/nbt.3001. Epub 2014 Aug 24. PMID: 25150839; PMCID: PMC4243706.

Nature Biotechnology 2014 Sep;32(9):926-32

overexpressed pathways underlying the MOA of the chemical.

We found that (i) the concordance between microarray and sequencing platforms for detecting the number of DEGs was positively correlated with the degree of the perturbation elicited by the treatment, (ii) RNA-seq

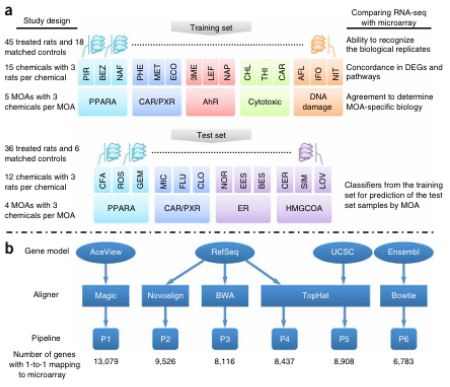


Figure 1 Overview of study design. (**a**) The study comprised a training set and a test set with the text on the left detailing the experimental design and the text on the right listing the key analyses conducted. Both microarray and RNA-seq were used to profile transcriptional responses induced by treatment of rats by each chemical; each is associated with a specific MOA. For each MOA there were three representative chemicals and three biological replicates per chemical. We evaluated cross-platform concordance at multiple levels: DEGs, mechanistic pathways and MOAs. To compare the predictive potential of RNA-seq and microarray as gene-expression biomarkers, we analyzed four MOAs by both platforms as a test set, two of the MOAs (PPARA and CAR/PXR) appear in the training set whereas the other two do not. (**b**) Overview of RNA-seq analysis pipelines.

nature biotechnology

ARTICLES

A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium

SEQC/MAQC-III Consortium*

We present primary results from the Sequencing Quality Control (SEQC) project, coordinated by the US Food and Drug Administration. Examining Illumina HiSeq, Life Technologies SOLiD and Roche 454 platforms at multiple laboratory sites using reference RNA samples with built-in controls, we assess RNA sequencing (RNA-seq) performance for junction discovery and differential expression profiling and compare it to microarray and quantitative PCR (qPCR) data using complementary metrics. At all sequencing depths, we discover unannotated exon-exon junctions, with >80% validated by qPCR. We find that measurements of relative expression are accurate and reproducible across sites and platforms if specific filters are used. In contrast, RNA-seq and microarrays do not provide accurate absolute measurements, and gene-specific biases are observed for all examined platforms, including qPCR. Measurement performance depends on the platform and data analysis pipeline, and variation is large for transcript-level profiling. The complete SEQC data sets, comprising >100 billion reads (10Tb), provide unique resources for evaluating RNA-seq analyses for clinical and regulatory settings.

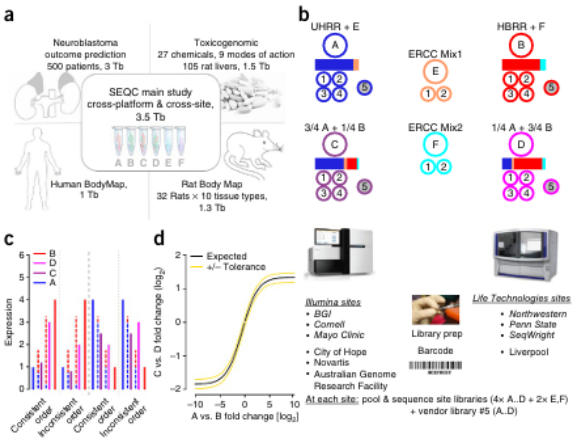


Figure 1 The SEQC (MAQC-III) project and experimental design. (**a**) Overview of projects. We report on a group of studies assessing different sequencing platforms in real-world use cases, including transcriptome annotation and other research applications, as well as clinical settings. This paper focuses on the results of a multicenter experiment with built-in ground truths. (**b**–**d**) Main study design. Similar to the MAQC-I benchmarks, we analyzed RNA samples A to D. Samples C and D were created by mixing the well-characterized samples A and B in 3:1 and 1:3 ratios, respectively. This allows tests for titration consistency (**c**) and the correct recovery of the known mixing ratios (**d**). Synthetic RNAs from ERCC were both added to samples A and B before mixing and also sequenced separately to assess dynamic range (samples E and F). Samples were distributed to independent sites for RNA-seq library construction and profiling by Illumina's HiSeq 2000 (three official + three unofficial sites) and Life Technologies' SOLiD 5500 (three official sites + one unofficial site). Unless mentioned otherwise, data show results from the three official sites (*italics*). In addition to the four replicate libraries each for samples A to D per site, for each platform, one vendor-prepared library A5…D5 was being sequenced at the official sites, giving a total of 120 libraries. At each site, every library has a unique bar-code sequence, and all libraries were pooled before sequencing, so each lane was sequencing the same material, allowing a study of lane-specific effects. To support a later assessment of gene models, we sequenced samples A and B by Roche 454 (3×, no replicates, see **Supplementary Notes**, section 2.5). (**c**) Schema illustrating tests for titration order consistency. Four examples are shown. The dashed lines represent the ideal mixture of samples A and B expected for samples D and C. (**d**) Schema illustrating a consistency test for recovering the expected sample mixing ratio. The yellow lines mark a 10% deviation from the expected response (black) for a perfect mixing ratio. Both tests (**c**) and (**d**) will reflect both systemic distortions (bias) and random variation (noise).

SEQC/MAQC-III Consortium. Su Z, Labaj PP, Li S, Thierry-Mieg J, Thierry-Mieg D, Shi W, **Wang C (project leader)**. et. al A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. doi: 10.1038/nbt.2957. Epub 2014 Aug 24. PMID: 25150838; PMCID: PMC4321899.

Nature Biotechnology 2014 Sep;32(9):903-14

Detecting and correcting systematic variation in large-scale RNA sequencing data

Sheng Li[1,2,11], Paweł P Łabaj[3,11], Paul Zumbo[1,2,11], Peter Sykacek[3], Wei Shi[4], Leming Shi[5], John Phan[6], Po-Yen Wu[6], May Wang[6], Charles Wang[7], Danielle Thierry-Mieg[8], Jean Thierry-Mieg[8], David P Kreil[3,9] & Christopher E Mason[1,2,10]

High-throughput RNA sequencing (RNA-seq) enables comprehensive scans of entire transcriptomes, but best practices for analyzing RNA-seq data have not been fully defined, particularly for data collected with multiple sequencing platforms or at multiple sites. Here we used standardized RNA samples with built-in controls to examine sources of error in large-scale RNA-seq studies and their impact on the detection of differentially expressed genes (DEGs). Analysis of variations in guanine-cytosine content, gene coverage, sequencing error rate and insert size allowed identification of decreased reproducibility across sites. Moreover, commonly used methods for normalization (cqn, EDASeq, RUV2, sva, PEER) varied in their ability to remove these systematic biases, depending on sample complexity and initial data quality. Normalization methods that combine data from genes across sites are strongly recommended to identify and remove site-specific effects and can substantially improve RNA-seq studies.

The deep sampling capabilities and single-base resolution of RNA-seq have led to its adoption for a variety of studies of the transcriptome, which include many inter-site and large-scale studies such as the ENCODE Project, GEUVADIS, GTEx, the Epigenomics Roadmap, the human Brainspan Project and the Nonhuman Primate Reference Transcriptome Resource. However, it is notable that RNA-seq, just like microarrays, has taken many years to emerge as a trusted and established method, as experiments can suffer from lack of principled experimental design, poor sample quality, inconsistent library preparation or platform-specific measurement biases[1,2]. Indeed, when microarrays started being used to identify biomarkers for drug toxicity and disease, the US Food and Drug Administration (FDA) recognized that an effort was needed to assure data quality and inter-site and inter-platform reproducibility and to this end established the MicroArray Quality Control (MAQC) Consortium[3]. Through the consortium, experimental standards and control RNA samples were developed, along with quality assurance guidelines and standardized microarray procedures[4]. Standards were also developed for data repositories (the minimum information about a microarray experiment, MIAME)[5], along with robust methods for analyzing microarray experiments from multiple sources[6]. These and other efforts have enabled the exploitation of the large publicly available microarray data sets and the subsequent deduction of important biological and clinical insights[7].

The success of MAQC motivated the development of similar guidelines and standards for high-throughput sequencing[8,9], in particular for RNA-seq[10,11], which led to the creation of the FDA Sequencing Quality Control (SEQC)/MAQC-III Consortium and the Association of Biomolecular Resource Facilities (ABRF) studies on Next-Generation Sequencing (NGS). Previous large-scale RNA-seq studies have focused on the variation between lanes and flow cells[12], and considerable progress has been made on reducing batch effects by normalizing GC content bias, fragment bias and the biases of isolation procedures[13–23]. So far, several RNA-seq data quality metrics have been developed[13,22,24,25], and surrogate variable analysis (sva)[26,27] has been applied to RNA-seq and microarray data from individual laboratories to improve expression measures[28]. Recently, a thorough, cross-site examination of Illumina RNA-seq data[29] demonstrated that "laboratory effects" strongly affect GC content and insert size of prepared RNA-seq libraries, and a method proposed to correct for them, probabilistic estimation of expression residuals (PEER)[30], was able to reduce artifacts without having an adverse impact on the detection of expression quantitative trait loci (eQTLs).

Yet, to date, there has been no systematic examination of the impact of site-specific bias in detecting DEGs, which is often the primary goal of an RNA-seq experiment. Moreover, there are various proposed means by which to correct for such biases, but the performance of several competing methods has not been systematically characterized. Here we used the controlled experimental design of the standardized SEQC/ABRF samples to test intra- and inter-site reproducibility,

[1]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York, USA. [2]The HRH Prince Alwaleed Bin Talal Bin Abdulaziz Alsaud Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York, USA. [3]Chair of Bioinformatics Research Group, Boku University Vienna, Vienna, Austria. [4]Department of Bioinformatics, WEHI, Melbourne, Australia. [5]State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, Schools of Life Sciences and Pharmacy, Fudan University, Shanghai, China. [6]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA. [7]Center for Genomics and Division of Microbiology & Molecular Genetics, School of Medicine, Loma Linda University, Loma Linda, California, USA. [8]National Center for Biotechnology Information (NCBI), Bethesda, Maryland, USA. [9]University of Warwick, Coventry, UK. [10]The Feil Family Brain and Mind Research Institute, New York, New York, USA. [11]These authors contributed equally to this work. Correspondence should be addressed to C.E.M. (chm2042@med.cornell.edu) or D.P.K. (david.kreil@boku.ac.at) or (d.kreil@warwick.ac.uk).

Received 17 April; accepted 28 July; published online 24 August 2014; doi:10.1038/nbt.3000

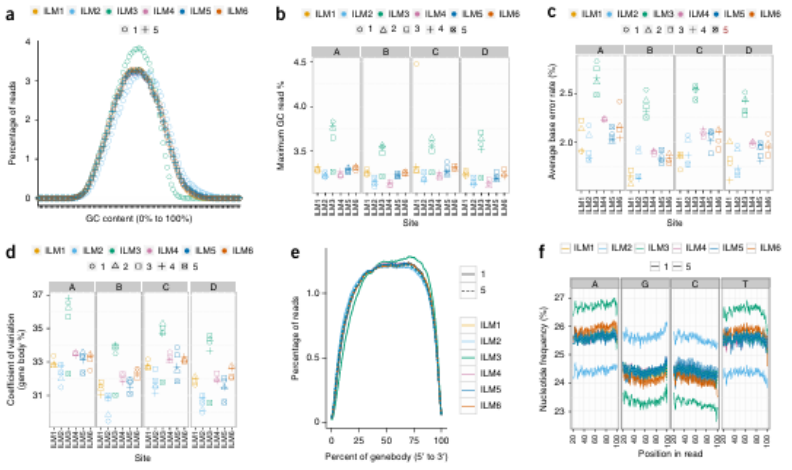


Figure 5 Examination of RNA-seq data quality identifies major sources of variation. (**a**) GC content distribution (sample A). Point shapes distinguish replicates (1: unfilled circle; 5: +). (**b**) The greatest percentage of reads contributing to some GC content bin (0–100%). A sample with more reads contributing to a particular GC content bin (%) indicates an abundance of reads with that particular GC content. (**c**) Average base error rate across all sequencing bases (*y* axis) across all sites (*x* axis). (**d**) Coefficient of variation of the percentage of gene body coverage (*y* axis), which is a measure of the evenness of coverage across all gene bodies for each site (*x* axis). (**e**) The percentage of reads that covers each nucleotide position of all of genes scaled to 100 bins, from 5′ UTR to 3′ UTR for sample A:1–5. Replicate 1 displayed site-dependent variation in gene body coverage for ILM3 (3′ bias), whereas replicate 5 showed similar gene body coverage regardless of where it was sequenced, suggesting that gene body coverage is influenced by library preparation. (**f**) Nucleotide frequency versus position for aligned reads. The percentage of each base was plotted as a function of the read length for each base (A, G, C, T) for two replicates (1, 5) for all sites. Replicate 1 displayed site-dependent base-composition frequencies, whereas replicate 5 showed similar base composition frequencies regardless of where it was sequenced, suggesting that base composition frequency is largely a result of library preparation. Only the 20th to the 100th bases are shown here; the full read range can be seen in **Supplementary Figure 4**. Vertical facets stand for sample A–D. Site information for ILM1–6 is color-coded. Replicates 1–4 were prepared and sequenced independently at each site, whereas replicate 5 was prepared at a single site and then sequenced at a subset of all sites. Point shapes distinguish replicates.

Li S, Łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, Phan J, Wu PY, Wang M, **Wang C,** Thierry-Mieg D, Thierry-Mieg J, Kreil DP, Mason CE. Detecting and correcting systematic variation in large-scale RNA sequencing data. doi: 10.1038/nbt.3000. Epub 2014 Aug 24. PMID: 25150837; PMCID: PMC4160374.

Nature Biotechnology 2014 Sep;32(9):888-95

nature COMMUNICATIONS

ARTICLE

Received 11 Aug 2014 | Accepted 1 Sep 2014 | Published 25 Sep 2014

DOI: 10.1038/ncomms6125

Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures

Sarah A. Munro[1,2], Steven P. Lund[1], P. Scott Pine[1,2], Hans Binder[3], Djork-Arné Clevert[4], Ana Conesa[5], Joaquin Dopazo[5,6], Mario Fasold[7], Sepp Hochreiter[4], Huixiao Hong[8], Nadereh Jafari[9], David P. Kreil[10,11], Paweł P. Łabaj[10], Sheng Li[12], Yang Liao[13,14], Simon M. Lin[15], Joseph Meehan[8], Christopher E. Mason[12], Javier Santoyo-Lopez[6,16], Robert A. Setterquist[17], Leming Shi[18], Wei Shi[13,19], Gordon K. Smyth[13,20], Nancy Stralis-Pavese[10], Zhenqiang Su[8,†], Weida Tong[8], Charles Wang[21], Jian Wang[22], Joshua Xu[8], Zhan Ye[23], Yong Yang[22], Ying Yu[18] & Marc Salit[1,2]

There is a critical need for standard approaches to assess, report and compare the technical performance of genome-scale differential gene expression experiments. Here we assess technical performance with a proposed standard 'dashboard' of metrics derived from analysis of external spike-in RNA control ratio mixtures. These control ratio mixtures with defined abundance ratios enable assessment of diagnostic performance of differentially expressed transcript lists, limit of detection of ratio (LODR) estimates and expression ratio variability and measurement bias. The performance metrics suite is applicable to analysis of a typical experiment, and here we also apply these metrics to evaluate technical performance among laboratories. An interlaboratory study using identical samples shared among 12 laboratories with three different measurement processes demonstrates generally consistent diagnostic power across 11 laboratories. Ratio measurement variability and bias are also comparable among laboratories for the same measurement process. We observe different biases for measurement processes using different mRNA-enrichment protocols.

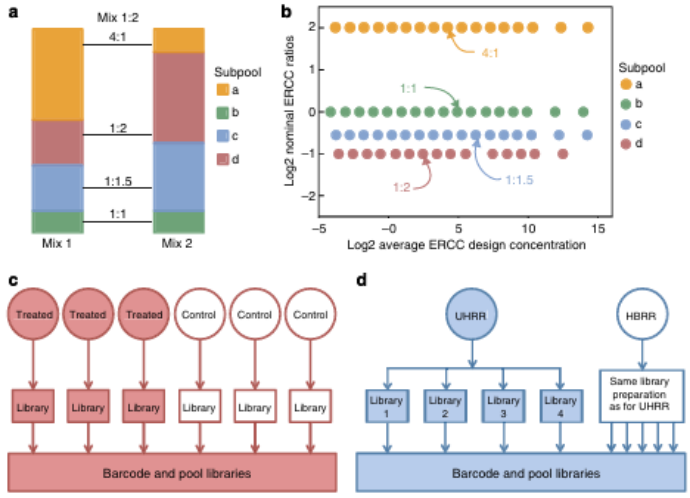


Figure 1 | Design of ERCC RNA control ratio mixtures and example experiments. (**a**) Two mixtures of the same 92 ERCC RNA transcripts are prepared such that four subpools with 23 transcripts per subpool are in four defined abundance ratios between the two mixtures. (**b**) Within each subpool the 23 controls (several points overlap) span a broad dynamic range of transcript concentrations. (**c**) In a typical single laboratory RNA-Seq experiment, biological replicates would be prepared for treatment and control samples. Rat toxicogenomics experimental samples represent this experimental design. (**d**) In the interlaboratory study design UHRR and HBRR samples have no biological replicates but have extensive technical replicates including multiple library preparation replicates that are analysed for the interlaboratory assessment of reproducibility instead of biological replicates.

Munro SA, Lund SP, Pine PS, Binder H, Clevert DA, Conesa A, Dopazo J, Fasold M, Hochreiter S, Hong H, Jafari N, Kreil DP, Łabaj PP, Li S, Liao Y, Lin SM, Meehan J, Mason CE, Santoyo-Lopez J, Setterquist RA, Shi L, Shi W, Smyth GK, Stralis-Pavese N, Su Z, Tong W, **Wang C**, Wang J, Xu J, Ye Z, Yang Y, Yu Y, Salit M. Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. doi: 10.1038/ncomms6125. PMID: 25254650.

Nature Communications 2014 Sep 25;5:5125