nature research

Corresponding author(s): Amit Verma

Last updated by author(s): Nov 2, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Targeted sequencing conducted on primary samples. All variants reported with details. No software used for data collection.
Data analysis	Targeted Sequencing Analysis NGS sequences were aligned to genome build hg19 using Burrows-Wheeler aligner and read quality and depth were determined using GATKv4. Single nucleotide variations were determined using Mutect and insertion/deletion mutations were determined using Pindel. Variants were excluded from analysis if there were present in ExAC/gnomAD databases at ≥1% minor allele frequency. Variants were analyzed with a minimum sequence depth of 250x and a variant allele frequency of 5% for single nucleotide variants and 10% for insertion/deletion mutations. Variants detected between 50-250x depth were analyzed base on a proprietary confidence interval. The median depth of coverage in our study was 839x with a mean coverage of 908x. Variants with VAFs above 40% were excluded as potential germline variants. Statistical Analyses We calculated prevalence of somatic mutations and associated 95% confidence intervals of WTC-exposed participants using binomial assumptions for 10-year age groups (i.e., 30-39; 40-49; 50-59; 60-69; 70-79; 80 and older). We then performed a multivariable logistic regression among all responders to evaluate the association of WTC exposure and CH. Models controlled for age at the time of blood collection (as an ordinal variable), race/ethnicity (non-Hispanic white, non-Hispanic black, non-Hispanic Asian, non-Hispanic other races, Hispanic) and sex. Then, we repeated this analysis excluding 52 WTC-exposed EMS workers to allow for appropriate comparison of firefighter cohorts. Finally, two additional analyses that controlled for all the aforementioned confounders as well as self-reported smoking status (ever vs. never) were conducted, first, among the entire WTC-exposed responder cohort and then, among WTC-exposed firefighters. Because IAFF firefighters were not asked smoking status, these two models were restricted to FDNY and Nashville firefighters only.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All sequencing data for first responders and control has been deposited to the European Variation Archive Project with Accession ID: PRJEB49193. https://www.ebi.ac.uk/eva/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	481 first responders and 255 non WTC exposed fireighters. Sample size for first responders based on availability of samples. The rates of CH were significantly different between the first responders and controls, justifying the sample sizes used in the study.
Data exclusions	No samples excluded from the study
Replication	Since each subject was sequenced once, there are no replicates. The sequencing was done using a rigorous clinical grade CLIA approved lab and platform.
Randomization	Not applicable as this is not a randomized therapeutic study.
Blinding	Not applicable as the first responders and controls were identified before sequencing. Sequncing results were blinded though as were reported as de-identified IDs.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☐	☒	Antibodies
☐	☒	Eukaryotic cell lines
☒	☐	Palaeontology and archaeology
☐	☒	Animals and other organisms
☐	☒	Human research participants
☒	☐	Clinical data
☒	☐	Dual use research of concern

Methods

n/a	Involved in the study	
☒	☐	ChIP-seq
☐	☒	Flow cytometry
☒	☐	MRI-based neuroimaging

Antibodies

Antibodies used

Antibody	Clone	Dilution	Supplier	Cat # L	Lot #
Gr1 - PE	RB6-8C5	1:100	Invitrogen	12-5931-82	2124558
CD4 - PE-Cy5	GK1.5	1:100	Invitrogen	15-0041-82	2298697
CD8α - PE-Cy5	53-6.7	1:100	Invitrogen	15-0081-82	2254274
B220 - PE-Cy5	RA3-6B2	1:100	Invitrogen	15-0452-82	2151468
CD19 - PE-Cy5	eBio1D3	1:100	Invitrogen	15-0193-82	2173323
Ter119 PE-Cy5	TER-119	1:100	Invitrogen	15-5921-82	1994119
Sca1 – BV421 (PB)	D7	1:100	BioLegend	108127	B266527
c-kit - APC	2B8	1:100	Invitrogen	17-1171-82	1990178
CD34 - FITC	RAM34	1:100	BD Pharminngen	560238	9322926

	FcγR II/III (CD16/32) – PE-Cy7 93 1:100 Invitrogen 25-0161-82 1994144
Validation	All commercial antibodies used very widely in numerous publications. all validated.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	GM03798 (wild type), Epstein–Barr virus-transformed lymphoblasts were obtained from Coriell Cell Repositories
Authentication	authenticated at Coriell with micro satellite analysis
Mycoplasma contamination	Negative by testing with mycoplasma kits
Commonly misidentified lines (See ICLAC register)	None

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Female WT C57BL/6 mice (The Jackson Laboratory) were used for all murine experiments
Wild animals	No wild animals were used for this work
Field-collected samples	The study did not have samples collected from the field
Ethics oversight	All murine experiments were approved by the New York University Grossman School of Medicine Institutional Animal Care and Use Committee protocol IA16-00447.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	481 first responders and 255 non WTC exposed firefighters, ages 30-100; 23 females
Recruitment	FDNY annual checkups: From December 2013 through October 2015, as part of these exams, whole blood samples were collected from 481 responders who were enrolled the study. The non-WTC-exposed comparison population included 52 patients recruited at the annual convention of the International Association of Firefighters (IAFF). An additional 203 Nashville firefighters were extracted from the Synthetic Derivative, a de-identified Electronic Health Record with paired biorepository (BioVU) of Vanderbilt University Medical Center, using natural language processing methodologies. For the selected 203 patient cohort, both firefighter occupation and smoking status were manually validated. Self selection bias is potential possible, though no clinical criteria was a factor in recruitment.
Ethics oversight	From December 2013 through October 2015, as part of these exams, whole blood samples were collected from 481 responders who were enrolled the study which was approved by the IRB of Albert Einstein College of Medicine (07-09-320) and who signed informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Murine bone marrow cells were used for FACS analysis to identify Kit+ Sca1+ Lineage-ve stem cells. The antibody cocktails used are listed below. KSL+ cells were sorted from WTC PM and control treated mice bone marrows and used for DNA isolation.

Instrument	FACS Aria
Software	FlowJo v10
Cell population abundance	LSK cells (1-2%) and LK cells (5-7%) as shown in Extended Data Figures
Gating strategy	Kit+ Sca1+ Lineage-ve stem cells gating stratgey shown in Extended Data Fig 7

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.