

# Using Big Data to Uncover Novel Insights into the Genetic Etiology of Cancer

*Mitchell J. Machiela, ScD MPH*

*Earl Stadtman Investigator*

*Integrative Tumor Epidemiology Branch Division  
of Cancer Epidemiology and Genetics National  
Cancer Institute*

# Big (Data) Question

How do we utilize genomic data to better understand cancer risk?

## Inherited

Rare, High-penetrant Mutations



Common Inherited Variants



## Acquired

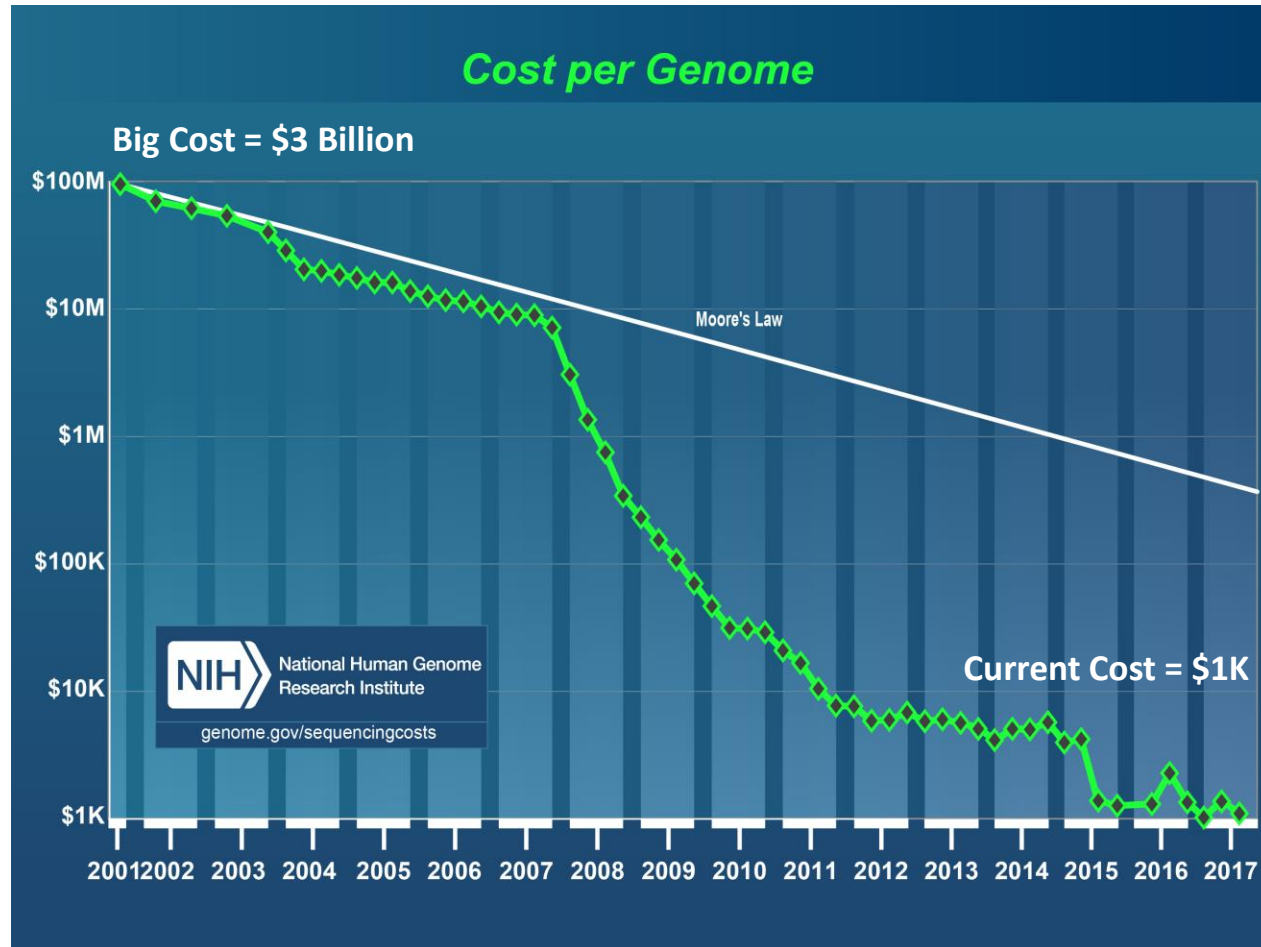


Tumor Genomes



Genetic Mosaicism

# Drastic Improvements and Cost Reductions in Sequencing/Genotyping Technology



Bloomberg the Company & Its Products | Bloomberg Anywhere Remote Login | Bloomberg Terminal Demo Request

Menu Search Bloomberg Sign In

## Prognosis

### A \$100 Genome Within Reach, Illumina CEO Asks If World Is Ready

By [Kristen V Brown](#)  
February 27, 2019, 2:04 PM EST

- ▶ In 2017, the company promised a \$100 genome within a decade
- ▶ CEO Francis deSouza says tech isn't the only thing in the way

Francis deSouza Photographer: Jeff Spicer/PA Wire via AP

# Inference Difficult from N=1

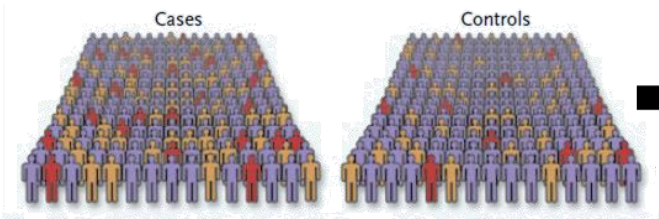
**Jeanne Calment**  
(1875-1997)



- Longest confirmed human lifespan on record (122 years, 164 days)
- Diet/Lifestyle:
  - Consumed over 2 pounds of chocolate a week
  - Drank copious amounts of port wine
  - Smoked for almost 100 years (>73,000 cigarettes!!)
- Never developed cancer

# Genome Wide Association Studies: GWAS 101

**Step 1:**  
Enroll and consent participants  
Collect and extract DNA



100,000s of Participants

**Step 2:**  
Hybridize and scan  
genotyping array



100,000s of Variants

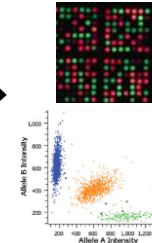
+

TopMed Reference (62,784 genomes)

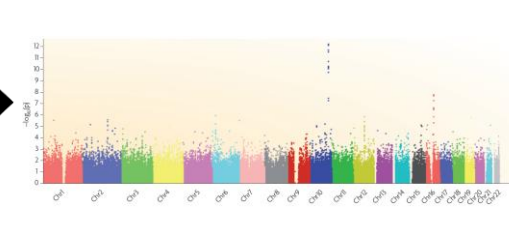
=

471,000,000 Variants

**Step 3:**  
Perform QC  
Call Genotypes



**Step 4:**  
Analyze each  
tagging SNP



10,000s Disease Associations

(<https://www.ebi.ac.uk/gwas/home>)

# Progress in Prostate Cancer

- Most common male non-skin cancer in the developed world
  - 174,650 new cases expected in US in 2019
  - 31,620 deaths expected in US in 2019
- Firmly established risk factors (2006 “Pre-GWAS era”)
  1. Increasing age
  2. Family history of prostate cancer
  3. Ancestry
    - African Americans 1.6x risk and 2.4x mortality

# Progress in Prostate Cancer

- 140,000 men across 50+ studies
  - 79,194 prostate cancer cases
  - 61,112 controls
- 160 germline susceptibility loci
  - highlight polygenic architecture
  - most with small effect sizes ( $OR < 1.1$ )



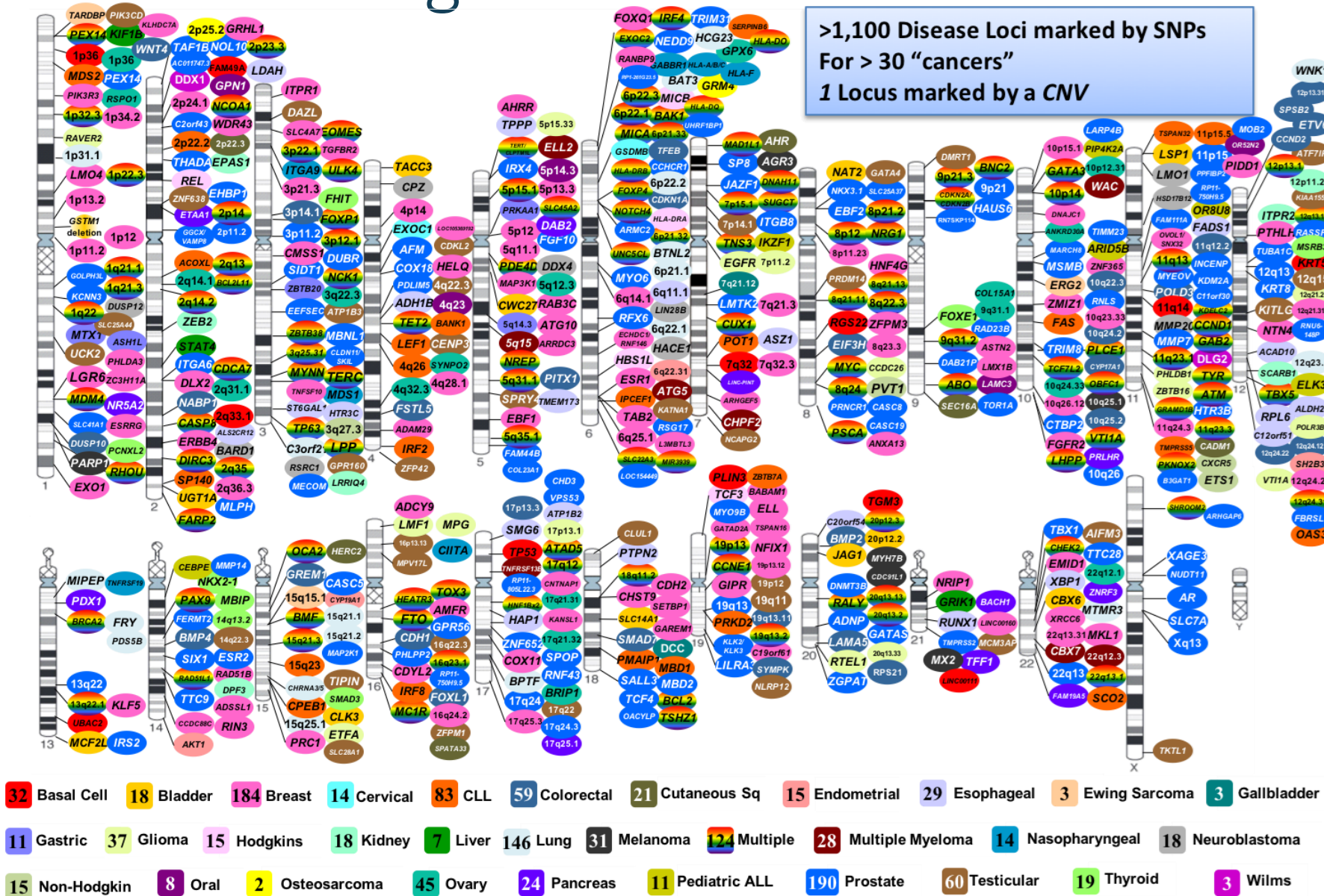


# Progress in Prostate Cancer

- Most common male non-skin cancer in the developed world
  - 174,650 new cases expected in US in 2019
  - 31,620 deaths expected in US in 2019
- Firmly established risk factors (2019)
  1. Increasing age
  2. Family history of prostate cancer
  3. Ancestry
  4. Approximately 160 germline susceptibility regions
    - new genetic loci for improved understanding of disease etiology
    - future value for individualized genetic risk prediction



# Progress in Cancer GWAS



# Progress in Cancer GWAS

Discovering the causes of cancer and the means of prevention

[DCEG Home](#)

[About DCEG](#)

[Our Research](#)

[Fellowships & Training](#)

[Tools & Resources](#)

[News & Events](#)

[Publications](#)

## Our Research

[Cancer Types](#)

[What We Study](#)

[Who We Study](#)

[How We Study](#)

[Active Clinical  
Studies](#)

[Public Health Impact  
of DCEG Research](#)

## Confluence project

[Print This Page](#)



The Confluence project will develop a large research resource by 2020 to uncover breast cancer genetics through genome-wide association studies (GWAS). The resource will include at least 300,000 breast cancer cases and 300,000 controls of different races/ethnicities. This will be accomplished by the confluence of existing GWAS and new genome-wide genotyping data to be generated through this project.

Broad scientific aims that can be addressed through this resource include:


1. To discover susceptibility loci and advance knowledge of etiology of breast cancer overall and by subtypes.
2. To develop polygenic risk scores and integrate them with known risk factors for personalized risk assessment for breast cancer overall and by subtypes.
3. To discover loci for breast cancer prognosis, long-term survival, response to treatment, and second breast cancer.

### Eligibility criteria

To be eligible to participate, studies with cases of *in situ* or invasive breast cancer (females or males) must have:

- Genome-wide genotyping data or germline DNA for genotyping, i.e.:
  - existing genome-wide genotyping data, or
  - germline DNA available for new genotyping, or
  - blood/buccal samples for germline DNA isolation and genotyping.
- Basic phenotype data (e.g. age at diagnosis, gender, family history of breast cancer)
- Appropriate ethics approval for genetic studies and data sharing

Please refer to the [Confluence study protocol \(pdf, 788 KB\)](#) for more details on Confluence and how studies can participate.

Please [complete the study inventory](#)  [Exit Disclaimer](#) if you are interested in participating in Confluence. This inventory is for planning purposes only and implies no commitment to participate.

Confluence is supported by NCI Intramural Research funds.

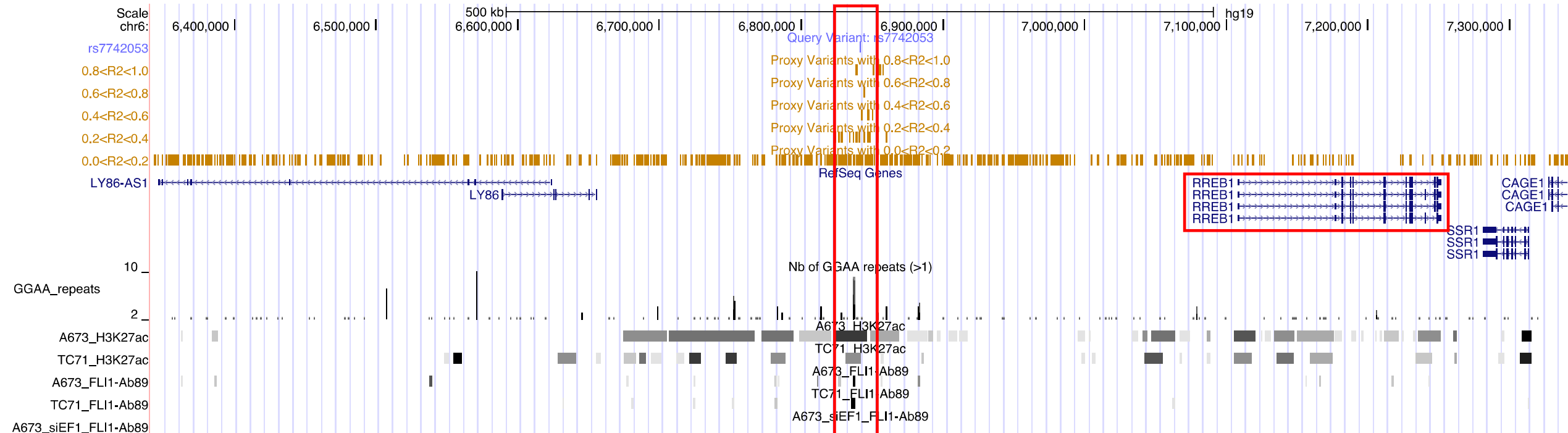
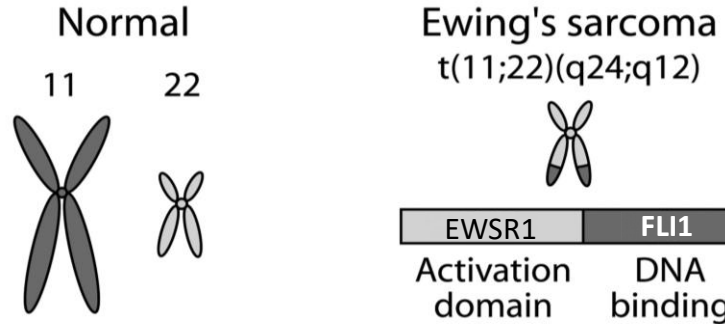
[Integrative Tumor Epidemiology Branch- Research Areas](#)

[DCEG Home](#) | [Contact DCEG](#) | [Policies](#) | [Accessibility](#) | [Viewing files](#) | [FOIA](#) | [DCEG Sitemap](#)

[U.S. Department of Health and Human Services](#) | [National Institutes of Health](#) | [National Cancer Institute](#) | [USA.gov](#)

NIH...Turning Discovery Into Health®

# Integrative Analysis: Ewing Sarcoma 6p25.1 Susceptibility Region



# Integrative Analysis



NATIONAL CANCER INSTITUTE

Division of Cancer Epidemiology & Genetics

[Contact Us](#) | [Staff Intranet](#) | [Sitemap](#)

Discovering the causes of cancer and the means of prevention

[DCEG Home](#)

[About DCEG](#)

[Our Research](#)

[Fellowships & Training](#)

[Tools & Resources](#)

[News & Events](#)

[Publications](#)

## Our Research

[Cancer Types](#)

[What We Study](#)

[Who We Study](#)

[How We Study](#)

[Active Clinical  
Studies](#)

[Public Health Impact  
of DCEG Research](#)

## Sherlock-lung: A Genomic Epidemiologic Study of Lung Cancer in Never Smokers

[Print This Page](#)

*Sherlock-lung* is a comprehensive study that aims to trace lung cancer etiology in never smokers by analyzing genomic data in tumor and surrounding lung tissue. Whole genome sequencing, whole transcriptome, and genome-wide methylation data will be analyzed to identify exogenous and endogenous processes involved in lung tumorigenesis. Analysis of the tumor microenvironment, clonal evolution, and circulating tumor DNA will be conducted in a subgroup of the cases. The molecular landscape will be integrated with histological and radiological features in order to develop a more refined classification of lung cancer in never smokers and provide insights into prognosis and treatment strategies.

*Sherlock-lung* will include 2,500 never smoker lung cancer patients, a subset (n=~500) with "special exposures," such as coal, radon, asbestos, air pollution, and microbial infection. The remaining ~2000 cases will come from the "general population," with unknown exposures to lung cancer risk factors.

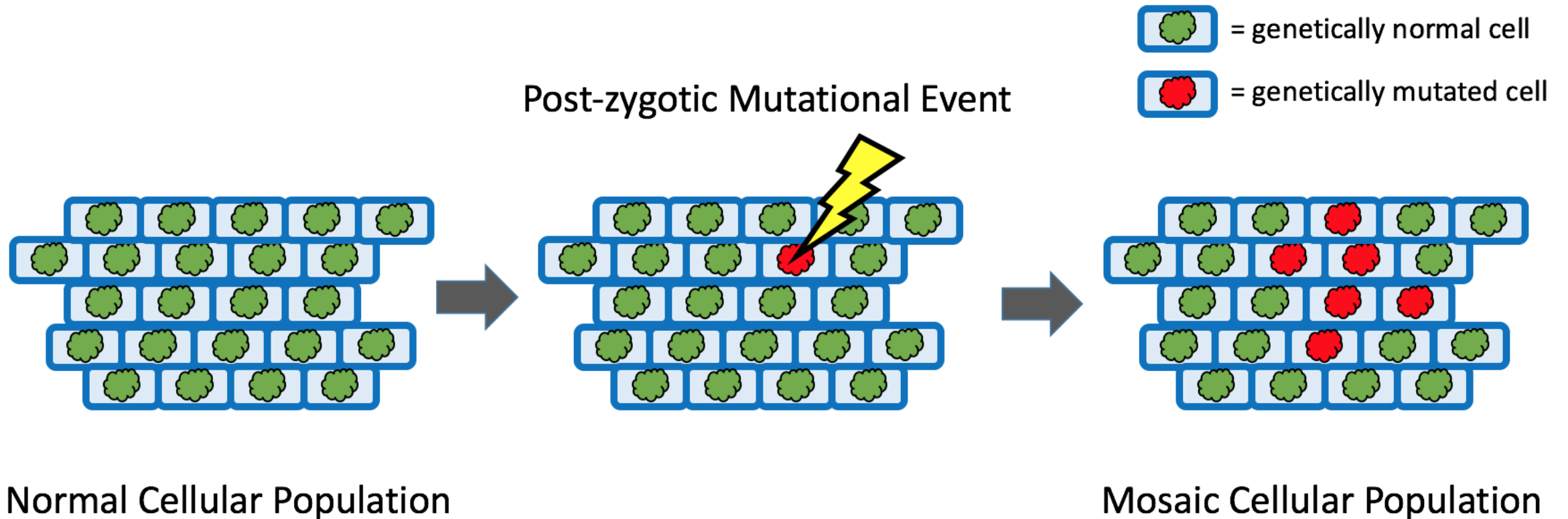
For more information about *Sherlock-lung*, please contact [Maria Teresa Landi, M.D., Ph.D.](#)

Integrative Tumor Epidemiology Branch - Research Areas

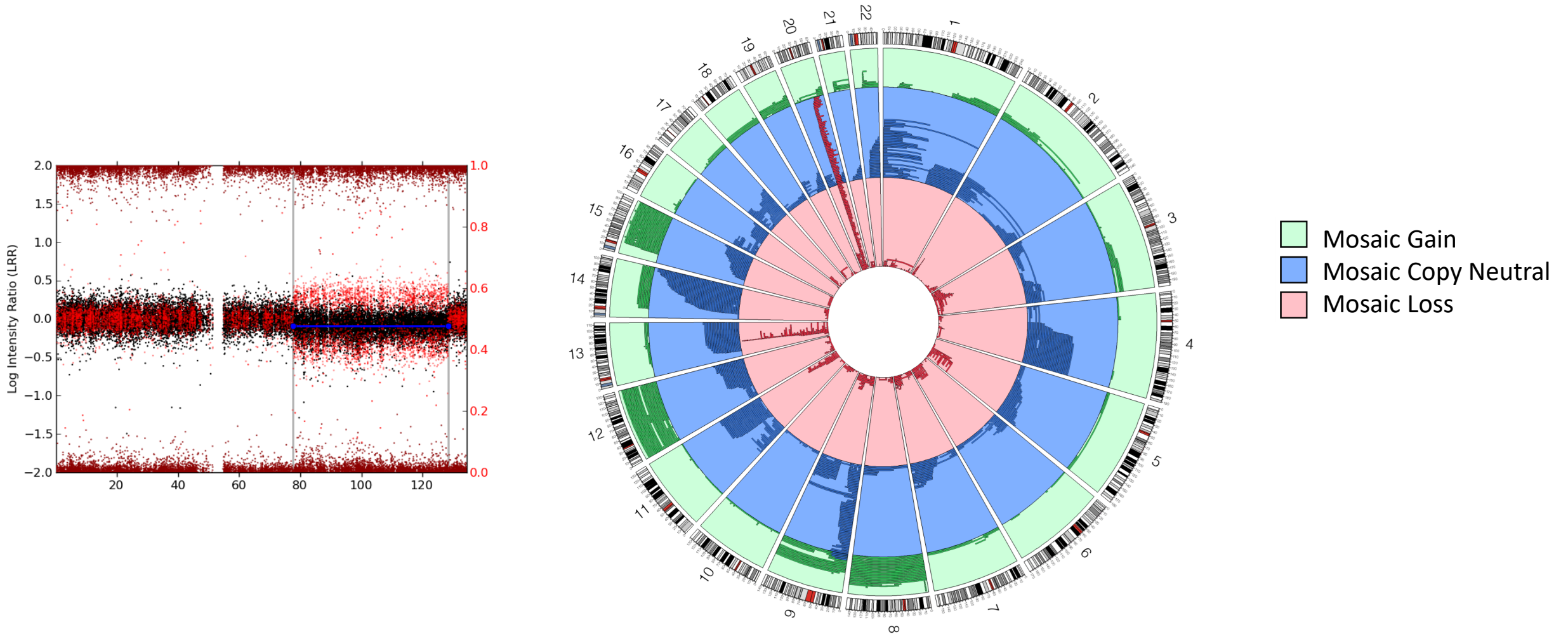


# Genetic Mosaicism

- **Definition:** the presence of an acquired mutation in a clonal population of cells that differs from the inherited genome

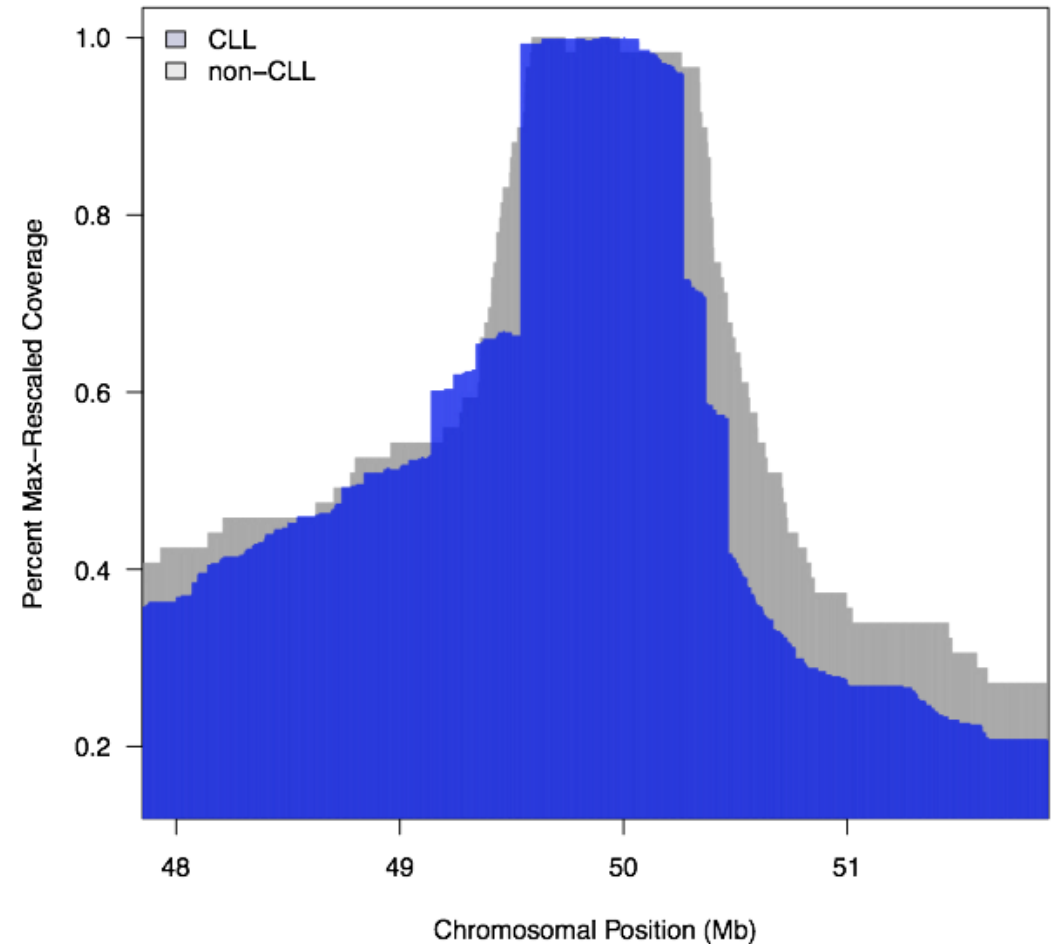


# Age-related Autosomal Mosaicism in 127K Healthy Participants



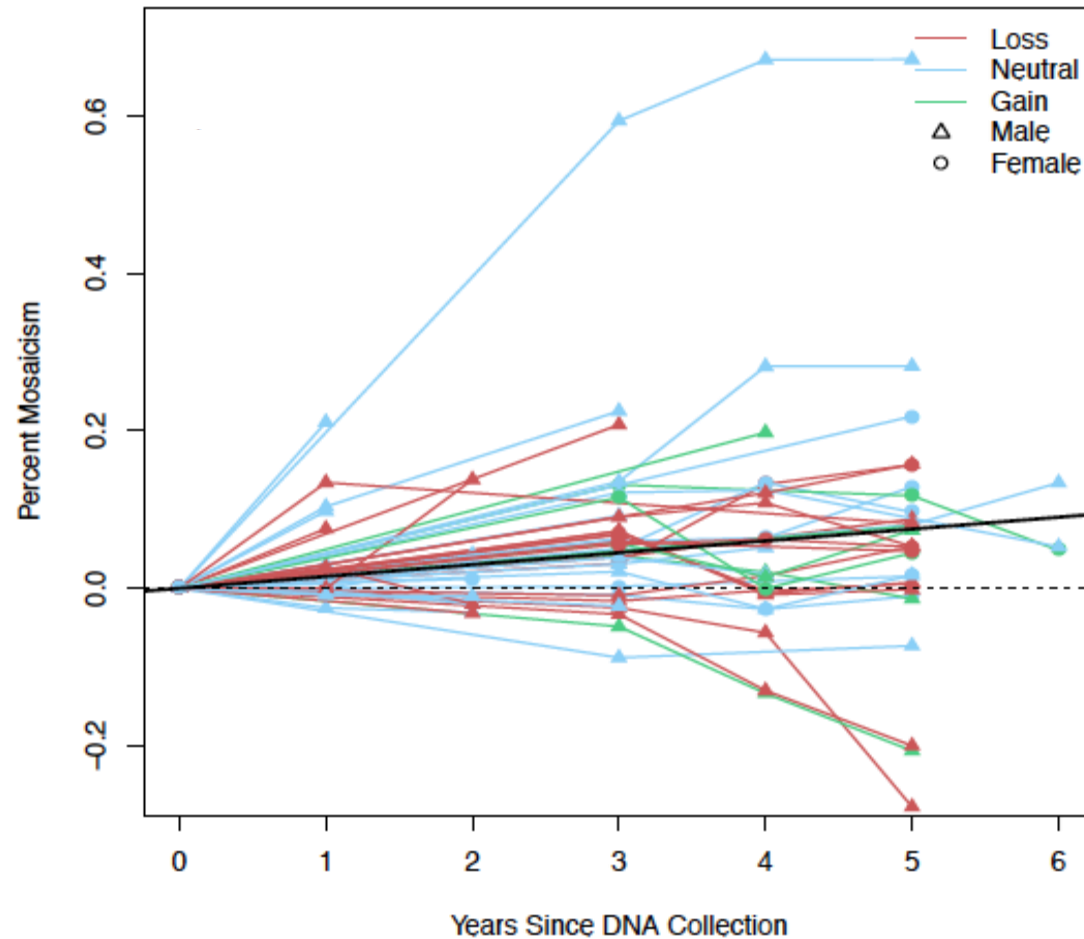
# Identification of Pre-leukemic Clones

- Mosaic 20q deletions cover same genomic footprint as del(20q) in myeloid malignancies
- Mosaic 13q14 deletions span the same genomic region as del(13q14) in MBL and CLL
- Frequencies of 20q and 13q14 deletions are higher than population rates of the respective disease
  - Not all individuals with mosaicism progress to disease
  - Having mosaicism in blood DNA increases hematologic cancer risk (OR=35)
  - Observed up to 15 years prior to diagnosis





# Longitudinal Data: Biology is a Dynamic Process



# Longitudinal Data

Discovering the causes of cancer and the means of prevention

[DCEG Home](#)

[About DCEG](#)

[Our Research](#)

[Fellowships & Training](#)

[Tools & Resources](#)

[News & Events](#)

[Publications](#)

## Our Research

[Cancer Types](#)

[What We Study](#)

[Who We Study](#)

[Cases and Controls](#)

[Cohorts](#)

[Families](#)

[How We Study](#)

[Active Clinical  
Studies](#)

[Public Health Impact  
of DCEG Research](#)

## Connect Study

[Print This Page](#)



The Connect study is a new prospective cohort of 200,000 adults in the United States designed to further investigate the etiology of cancer and its outcomes, which may inform new approaches in precision prevention and early detection. The new cohort will capitalize on research innovations to advance the field of cancer epidemiology and prevention including:

1. New technologies (e.g., tracking and sensors to measure behavior and environment);
2. Large-scale analyses of the genome, epigenome, transcriptome, proteome, metabolome, microbiome;
3. Molecular profiling of tumors and precursor lesions to study the natural history of cancer and its etiologic heterogeneity.

### Overview of Study Setting and Design

The Connect study will be conducted within a set of integrated health care systems, with electronic medical records (EMRs), a passive follow-up system that is both cost effective and thorough. Consented participants ages 40-65 with no history of invasive cancer other than non-melanoma skin cancer from participating health systems will complete an online questionnaire at baseline and periodically throughout the duration of follow-up. Passive follow-up via tumor registries and EMRs will provide outcome information for cancers and their precursors. Blood, urine and saliva samples will be collected at baseline and repeatedly during follow-up in local clinics. Additional biological specimens including fecal and tissue specimens will be collected. This state-of-the-art cohort will be built with an efficient, flexible and integrated infrastructure that makes the most of modern interoperability standards in order to serve as a research workhorse for future generations of scientists at the NCI and across the extramural research community.

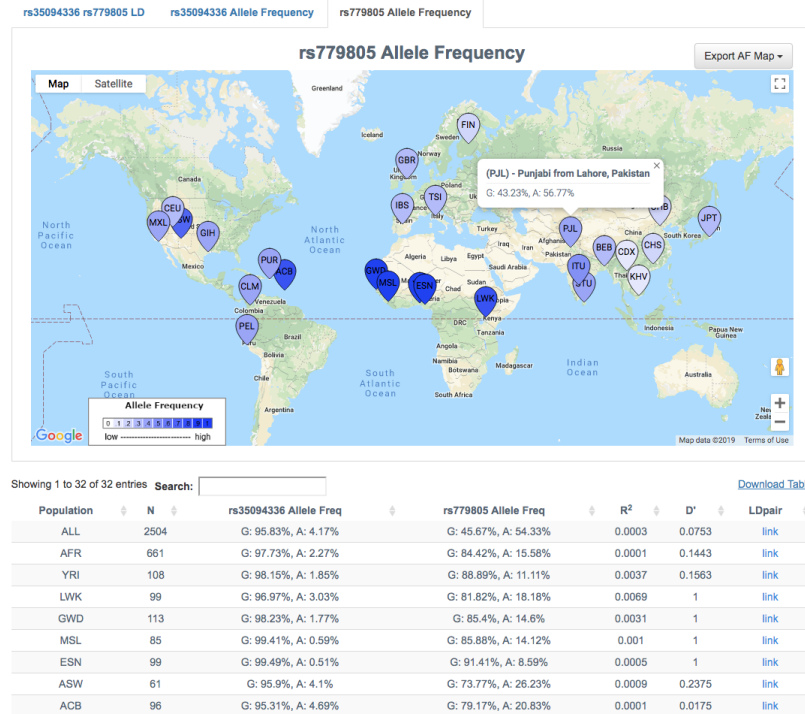


[Enlarge](#)

The Connect study is supported by the [NIH Intramural Research Program](#).

# Tools to Harness Big Data

## LDlink



[ldlink.nci.nih.gov](https://ldlink.nci.nih.gov)

## AuthorArranger

Conquer journal title pages in seconds

**Author Format** **Affiliation Format**

**Fields (drag to reorder)**

Field	Format	Options	Action
Title	(None)	<input checked="" type="checkbox"/> Add Period	x
First	First Name	<input type="checkbox"/> Abbreviate <input type="checkbox"/> Add Period	x
Middle	Middle Name	<input type="checkbox"/> Abbreviate <input type="checkbox"/> Add Period	x
Last	Last Name	<input type="checkbox"/> Abbreviate <input type="checkbox"/> Add Period	x
Degree	(None)	<input type="checkbox"/> Add Comma <input type="checkbox"/> Add Period	x
Other	(None)	<input type="checkbox"/> Add Comma <input type="checkbox"/> Add Period	x

**Email Column**

Field	Format	Options	Action
Email	Email	<input checked="" type="checkbox"/> Add Semicolon	x

**Author Affiliation**

☒ Superscript ☐ Inline ☐ Subscript

**Author Separator**

☒ Comma ☐ Newline ☐ Semicolon ☐ Other

**Preview** **Reorder** **Email** **Download .docx**

Lawrie Wheeler<sup>54</sup>, Richard A. Sturm<sup>35</sup>, Amy Hutchinson<sup>1,92</sup>, Kristine Jones<sup>1,92</sup>, Michael Malasky<sup>1,92</sup>, Aurelie Vogt<sup>1,92</sup>, Weiyin Zhou<sup>1,92</sup>, Karen A. Pooley<sup>93</sup>, David E. Elder<sup>94</sup>, Jiali Han<sup>54</sup>, Belynda Hicks<sup>1,92</sup>, Nicholas K. Hayward<sup>95</sup>, Peter A. Kanetsky<sup>96</sup>, Chad Brummett<sup>97</sup>, Grant W. Montgomery<sup>98</sup>, Catherine M Olsen<sup>99</sup>, Caroline Hayward<sup>100</sup>, Alison M. Dunning<sup>101</sup>, Nicholas G. Martin<sup>48</sup>, Evangelos Evangelou<sup>102,103</sup>, Graham J. Mann<sup>104,105</sup>, Paul D. P. Pharoah<sup>101</sup>, Douglas F. Easton<sup>93</sup>, Jennifer H. Barrett<sup>10</sup>, Anne E. Cust<sup>106,107</sup>, Goncalo Abecasis<sup>108</sup>, David L. Duffy<sup>48,109</sup>, David C. Whiteman<sup>99</sup>, Helen Gogas<sup>110</sup>, Arcangela De Nicolò<sup>111</sup>, Margaret A. Tucker<sup>1</sup>, Julia A. Newton Bishop<sup>53</sup>, Ketty Peris<sup>22,23</sup>, Stephen J. Chanock<sup>1</sup>, Kevin M. Brown<sup>1</sup>, Florence Demenais<sup>6</sup>, Susana Puig<sup>15</sup>, Eduardo Nagore<sup>19</sup>, Jianxin Shi<sup>1</sup>, Mark M. Iles<sup>10</sup>, Matthew H. Law<sup>3</sup>, GenoMEL Consortium<sup>112</sup>, Q-MEGA and QTWIN Investigators<sup>112</sup>, ATHENS Melanoma Study Group<sup>112</sup>, 23andMe<sup>112</sup>, The SDH Study Group<sup>112</sup>, Essen-Heidelberg Investigators<sup>112</sup>, AMFS Investigators<sup>112</sup>, MelaNostrum Consortium<sup>112</sup>, Richard Scoyler<sup>113,114,115</sup>, Georgina Long<sup>116</sup>

<sup>1</sup>Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA,  
<sup>2</sup>Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK & Leeds Institute for Data Analytics, University of Leeds, Leeds, UK,  
<sup>3</sup>Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Australia,  
<sup>4</sup>Department of Dermatology, Andreas Syggros Hospital, Medical School, National and Kapodistrian University of Athens, Athens, Greece,  
<sup>5</sup>Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA

[authorarranger.nci.nih.gov](https://authorarranger.nci.nih.gov)

# Key Points on Big Data

- Technologic improvements accelerate acquisition of big data
- Large samples enable discovery of novel etiologic insights
- Integrative analyses provide clues to biologic processes
- Importance of longitudinal data often overlooked
- Independent replication is essential



**mitchell.machiela@nih.gov**



**NATIONAL  
CANCER  
INSTITUTE**

**DCEG is seeking talented fellows: <https://dceg.cancer.gov/fellowship-training>**

**[www.cancer.gov](http://www.cancer.gov)**