An Introduction to HipSci data

The Human Induced Pluripotent Stem Cells Initiative (HipSci) is generating a large, high-quality reference panel of human IPSC lines for the research community. These lines are created from tissue donations from both healthy volunteers and patients from particular rare disease communities.

This document and its associated video describes the data generated by the HipSci project and how user-access to the data is managed.

Each line generated by HipSci is extensively characterised, with genetic, proteomic and phenotypic data that is freely available for use by the wider research community.

The HipSci project releases both its quality control data (QC) and its assay data to the community so everyone can use these datasets in research.

Quality Control Data

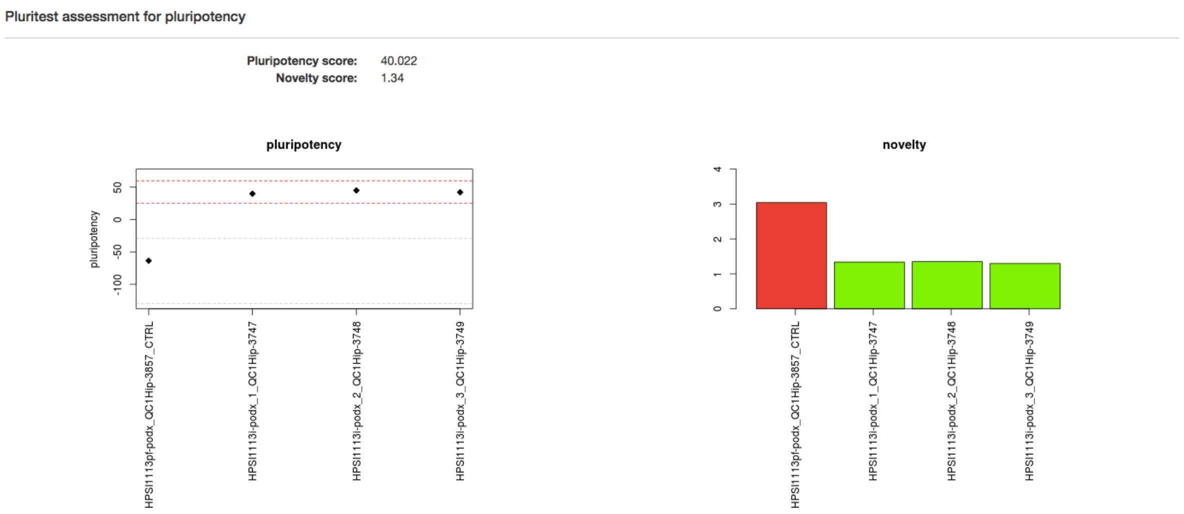
The HipSci project runs two quality control assays (QC) on two candidate cell lines (one line only when two lines did not make it through the pipeline) from each donor, and on the somatic cells that were used to derive it (fibroblasts or blood). From this QC, HipSci assesses pluripotency and genetic stability of the lines and selects suitable cell lines for banking. For cell lines derived during the earliest phases of the project, three lines per donor were picked as candidates, and in some cases two of these were selected for banking.

Pluripotency Assessment (Expression array)

The candidate iPSC lines and somatic cells are both assayed using an expression array such as the HumanHT-12 v4 Expression BeadChip Kit from Illumina. The results of these assays are then used by Pluritest[1], a tool which compares to expression data from a training set of 450 lines to assess the level of expression of pluripotency markers in the given sample.

You can see the Pluritest results for a particular cell line on its cell line page such as http://www.hipsci.org/lines/#/lines/HPSI1113i-podx_1. The two charts show the calculated pluripotency score and the novelty score, respectively, for the iPSC clones assayed as well as the control somatic cell lines.

[1] 1 A bioinformatic assay for pluripotency in human cells, F Muller et al, *Nature Methods* **8**, 315–317 (2011) doi:10.1038/nmeth.1580



Pluritest assessment for pluripotency

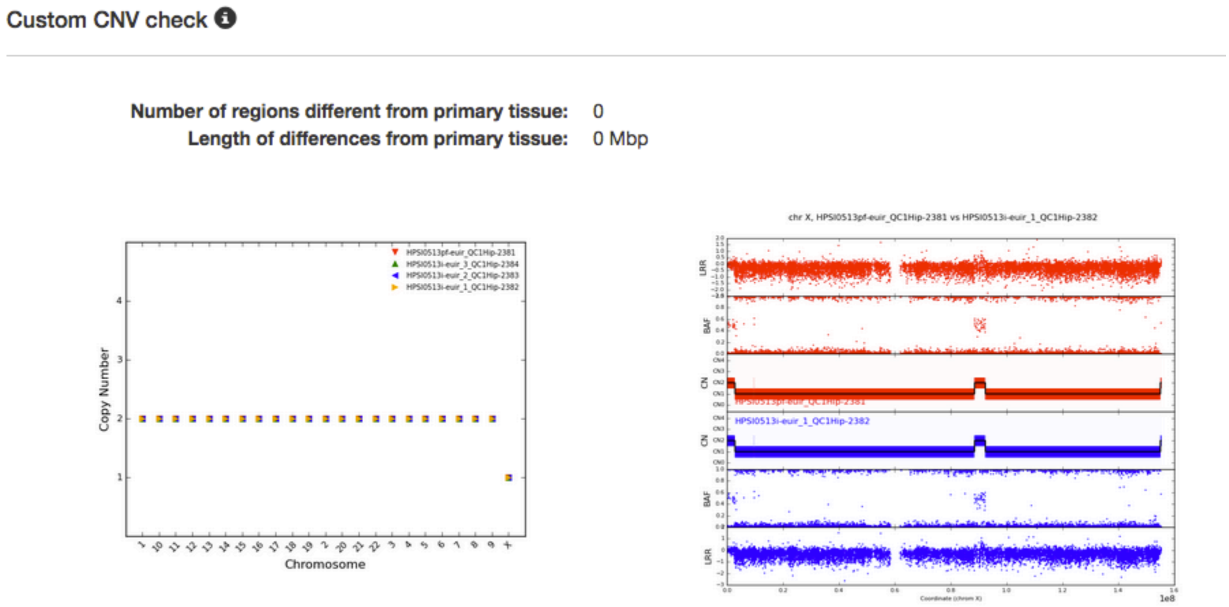
Pluripotency score: 40.022
Novelty score: 1.34

pluripotency
pluripotency
50
0
-50
-100
HPSI1113pf-podx_QC1Hip-3857_CTRL
HPSI1113i-podx_1_QC1Hip-3747
HPSI1113i-podx_2_QC1Hip-3748
HPSI1113i-podx_3_QC1Hip-3749
novelty
4
3
2
1
0
HPSI1113pf-podx_QC1Hip-3857_CTRL
HPSI1113i-podx_1_QC1Hip-3747
HPSI1113i-podx_2_QC1Hip-3748
HPSI1113i-podx_3_QC1Hip-3749

Genetic Stability Assessment (Genotyping by array)

The iPSC clones and somatic cells are both genotyped using a beadchip array from Illumina. Like for the pluripotency assessment, the results from the iPSC lines are compared with the results from the somatic cells that were assayed. The comparison is conducted using an algorithm[2] written specifically for the HipSci project.

The custom CNV results are visible on each cell line page, such as http://www.hipsci.org/lines/#/lines/HPSI0513i-euir_1 and show if there are any copy number aberrations between the somatic derived control data and the given iPSC lines. The first plot shows a summary of the analysis across each chromosome; the subsequent plots show any aberrant region in more detail.



Custom CNV check

Number of regions different from primary tissue: 0
Length of differences from primary tissue: 0 Mbp

HPSI0513pf-euir_QC1Hip-2381
HPSI0513i-euir_3_QC1Hip-2384
HPSI0513i-euir_2_QC1Hip-2383
HPSI0513i-euir_1_QC1Hip-2382
Copy Number
4
3
2
1
Chromosome
1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
3
4
5
6
7
8
9
X
chr X, HPSI0513pf-euir_QC1Hip-2381 vs HPSI0513i-euir_1_QC1Hip-2382
LRR
BAF
CN
HPSI0513pf-euir_QC1Hip-2381
HPSI0513i-euir_1_QC1Hip-2382
CN
BAF
LRR
Coordinate (chrom X)
1e8

[2] A Method for Checking Genomic Integrity in Cultured Cell Lines from SNP Genotyping Data, P Danecek et al, PLoS One. 2016 May 13;11(5):e0155014. doi: 10.1371/journal.pone.0155014. eCollection 2016.

Characterisation Assays

Based on the QC results, particular iPSC lines are selected to be banked in a public cell bank These lines are then subject to several more genomic, proteomic and phenotyping assays.

These assays are conducted in three different institutes from the HipSci consortium. The Wellcome Trust Sanger Institute conducts all the genomic assays; Kings College London generates the cellular phenotyping data for the project and the University of Dundee conducts all the proteomic assays.

Whole exome sequencing and whole genome sequencing

Whole exome sequencing is performed on all iPSC lines selected for banking in order to genotype those lines in the exonic regions of the genome. A small number of lines were also selected for whole genome sequencing.

Furthermore, we have sequenced ~250 of the somatic cell lines from which the iPSCs were derived. These somatic cell lines are all from the healthy volunteer cohort.

RNA-Seq

RNA-Seq is performed on all selected iPSC lines. This provides the consortium with data to study phenomena such as gene expression levels, alternative gene-spliced transcripts, and gene fusion events.

Methylation Array

Methylation profiling by array is used to probe the methylation pattern of DNA, which is a suppressor of gene activity. Methylation is used in HipSci’s assessment of variability in the pluripotent phenotype, and its dependence on genetic and technical factors.

Cellular Phenotyping

The cellular phenotyping project at Kings College London, is evaluating how iPSCs respond to chemical, physical and biological stimuli. A high-content platform uses novel assays and artificial stimuli to analyse iPSC behaviour in different microenvironments. These data contribute to HipSci’s research into the dependence of phenotypic variance on genetic and epigenetic variance.

Proteomic Mass Spectometry

HipSci's proteomics data is generated at the University of Dundee. Mass spectrometry is used to assay the proteome of the HipSci iPS cell lines. Proteomics can be used to explore, protein modification, expression, movement and interactions in metabolic pathways.

Management of data access

HipSci puts all of its data into the public domain, to share it with researchers worldwide. Once in the public domain, data sets are classed as either “managed access”, meaning users must register to obtain access; or “open access”, meaning any user can download the data immediately without registering.

Managed access (M)

- The ethical consent agreement of some HipSci donors authorises release of individually unique data for specific research use to bona fide researchers.
- Cell lines and data are marked as 'Managed access' or 'M' in HipSci’s cell line and data browser, if the donor's individually unique data are bound by these restrictions.
- All managed access data are stored in the EGA archive.
- For access to these data, researchers must apply for access to the data via WTSI's Electronic Data Access Mechanism.

Open access (O)

- Open access donors authorize the release of individually unique data to all parties, with no requirement to satisfy any data access restrictions.
- Furthermore, data from *any* donor is classed as open access if it is not individually unique; i.e. does not contain genotype information. This includes proteomics mass spectrometry, and cellular phenotyping measurements.
- These cell lines are marked as 'Open access' or 'O' in HipSci’s cell line and data browser, and there are links to download the data directly.
- The Y chromosome is excluded from open access data before it is made publicly available. This is to protect and prevent the identification of our donors.

Where to get help

If you are struggling to find the information you need regarding our cell lines and data, or have another general enquiry, you can contact us at hipsci@ebi.ac.uk

Acknowledgements

Thanks to Ian Streeter, Laura Clarke and Reena Halai for producing this document.