Title:

Somatic mutations reveal asymmetric cellular dynamics in the early human embryo

Author list

Young Seok Ju[1,2], Inigo Martincorena[1], Moritz Gerstung[1,3], Mia Petljak[1], Ludmil B Alexandrov[1,4], Raheleh Rahbari[5], David C Wedge[1], Helen R Davies[1], Manasa Ramakrishna[1], Anthony Fullam[1], Sancha Martin[1], Christopher Alder[1], Nikita Patel[1], Steve Gamble[1], Sarah O'Meara[1], Dilip D Giri[6], Torril Sauer[7], Sarah E Pinder[8], Åke Borg[9,10,11], Henk Stunnenberg[12], Marc van de Vijver[13], Benita K.T. Tan[14], Carlos Caldas[15], Andrew Tutt[16], Naoto T Ueno[17], Laura J van't Veer[18], John W. M. Martens[19], Christos Sotiriou[20], Stian Knappskog[21,22], Paul N. Span[23], Sunil R. Lakhani[24,25,26], Jórunn Erla Eyfjörd[27], Anne-Lise Børresen-Dale[28,29], Andrea Richardson[30,31], Alastair M. Thompson[32], Alain Viari[33], Matthew E Hurles[5], Serena Nik-Zainal[1], Peter J Campbell[1] and Michael R Stratton[1]‡

Affiliations

1. Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK.
2. Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Republic of Korea.
3. European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK.
4. Theoretical Biology and Biophysics (T-6), Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA.
5. Genomic Mutation and Genetic Disease, Wellcome Trust Sanger Institute, Hinxton, UK.
6. Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, USA.
7. Institute of Clinical Medicine, Campus at Akershus University Hospital, University of Oslo, Lørenskog, Norway.
8. King's Health Partners Cancer Biobank, Guy's Hospital, King's College London School of Medicine, London, UK.
9. BioCare, Strategic Cancer Research Program, Lund, Sweden.
10. CREATE Health, Strategic Centre for Translational Cancer Research, Lund, Sweden.
11. Department of Oncology and Pathology, Lund University Cancer Center, Lund, Sweden.
12. Radboud University Medical Center, Nijmegen, The Netherlands.
13. Department of Pathology, Academic Medical Center, Amsterdam, The Netherlands.
14. Department of General Surgery, Singapore General Hospital, Singapore.
15. Cancer Research UK (CRUK) Cambridge Institute, University of Cambridge, Cambridge, UK.
16. Breakthrough Breast Cancer Research Unit, Research Oncology, King's College London, Guy's Hospital, London SE1 9RT, UK.
17. Department of Breast Medical Oncology, MD Anderson Cancer Center, Houston, Texas, USA
18. Department of Laboratory Medicine, Helen Diller Family Comprehensive Cancer Center, University of California, San Francisco, USA.
19. Department of Medical Oncology, Erasmus MC Cancer Institute, Erasmus University Medical Center, Rotterdam, Netherlands.
20. Institut Jules Bordet, Brussels, Belgium
21. Section of Oncology, Department of Clinical Science, University of Bergen, Norway.
22. Department of Oncology, Haukeland University Hospital, Bergen, Norway.
23. Department of Radiation Oncology and Department of Laboratory Medicine, Radboud University Medical Center, Nijmegen, Netherlands.
24. University of Queensland, School of Medicine, Brisbane, Australia.
25. Pathology Queensland, Royal Brisbane and Women's s Hospital, Brisbane, Australia.
26. University of Queensland, UQ Centre for Clinical Research, Brisbane, Australia.
27. Cancer Research Laboratory, University of Iceland, Reykjavik, Iceland.
28. Department of Genetics, Institute for Cancer Research, Oslo University Hospital, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway
29. The K.G. Jebsen Center for Breast Cancer Research, Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Norway
30. Dana-Faber Cancer Institute, 450 Brookline Avenue, Boston, MA 02215, USA.
31. Brigham and Women's Hospital, Harvard Medical School, 75 Francis St, Boston, MA 02115, USA.

32. Department of Breast Surgical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
33. Plateforme Gilles Thomas - Synergie Lyon Cancer, Centre Léon Bérard, Lyon Cedex 08, FRANCE.

‡ **Corresponding author**

Michael R Stratton

Director, Wellcome Trust Sanger Institute,
Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom
e-mail: mrs@sanger.ac.uk
Tel: +44 (0) 1223 494757

First paragraph

Somatic cells acquire mutations throughout the course of an individual's life. Mutations occurring early in embryogenesis will often be present in a substantial proportion of, but not all, cells in the post-natal human and thus have particular characteristics and impact[1]. Depending upon their location in the genome and the proportion of cells they are present in, these mosaic mutations can cause a wide range of genetic disease syndromes[2] and predispose to cancer[3,4]. They have a high chance of being transmitted to offspring as *de novo* germline mutations and, in principle, can provide insights into early human embryonic cell lineages and their contributions to adult tissues[5]. Although it is known that gross chromosomal abnormalities are remarkably common in early human embryos[6] our understanding of early embryonic somatic mutations is very limited. Here, we use whole genome sequences of adult normal blood from 241 individuals to identify 163 early embryonic mutations. We estimate that approximately three base substitution mutations occur per cell per cell-doubling in early human embryogenesis and these are mainly attributable to two known mutational signatures[7]. We used the mutations to reconstruct developmental lineages of adult cells and demonstrate that the two daughter cells of many early embryonic cell doublings contribute asymmetrically to adult tissues at an approximately 2:1 ratio. This study therefore provides insights into the mutation rates, the mutational processes and the developmental outcomes of cell dynamics operative during early human embryogenesis.

Main text

Somatic point mutations of early embryonic derivation that are present in a substantial proportion of adult cells are detectable by standard DNA sequencing approaches. They can be distinguished from inherited single nucleotide polymorphisms (SNPs) as they will generally show lower variant allele fractions (VAFs). For example, early somatic mutations arising in one of the two daughter cells of the fertilized egg will show VAFs of ~25% in an adult tissue (**Fig. 1a**), compared to ~50% for inherited heterozygous polymorphisms, if the two cells have contributed equally to the adult tissue analysed[8]. Therefore, to identify early embryonic base substitutions, we analysed whole-genome sequences of blood samples from 279 individuals with breast cancer, in the first instance seeking mutations with VAFs ranging from 10% to 35%. However, with sequencing coverage of these samples at a median of 32-fold (**Supplementary Table 1**) a substantial number (50,000-100,000) of inherited heterozygous SNPs would also be expected, by chance, to show VAFs less than 35% thus interfering with specific detection of early embryonic mutations. To address this problem, we examined the relationship of low VAF substitutions to nearby (<500bp distant) heterozygous germline SNPs. Early embryonic mutations will be found on just a fraction of the sequencing reads that also carry a particular allele of a nearby heterozygous germline polymorphism (**Fig. 1b**). By contrast, an inherited SNP that by chance has a low VAF will be found on all such reads. Given the presence of approximately two million heterozygous SNPs in an individual human genome, a significant number of low VAF base substitutions can be interrogated in this manner. After further filtering and subsequent experimental validation by ultrahigh-depth targeted sequencing (median read-depth=22,000x), we identified 605 somatic base substitutions with highly accurate VAF estimates that appeared to be present in only a proportion of adult blood cells (Methods).

Mutations present in a subset of white blood cells can also reflect the presence of neoplastic clonal expansions arising from adult haematopoietic stem cells during the course of life[9-11]. However, blood samples in which neoplastic clones constitute a substantial proportion of the cells present will often display large numbers of low VAF mutations, because mutations present in the most recent common ancestor (MRCA) cell of the neoplastic clone have

usually been acquired over decades, rather than a few days for mutations of early embryonic origin (**Extended Data Fig. 1)**. Indeed, the distribution of the number of mosaic mutations in the 279 samples indicated the presence of an outlier group of 31 samples harbouring more mutations (n≥5) than expected by chance, compatible with the presence of neoplastic clones in these samples (**Fig. 1c**). Several additional features support the proposition that the low VAF mutations in these blood samples were predominantly from neoplastic clones (**Extended Data Fig. 1)**. These include: their absence from the breast cancers from the same individuals (**Figs. 1c-1e**); the frequent presence of known driver mutations for haematological neoplasms; the median age of individuals carrying these clones being twelve years higher than the rest of the cases (64 vs. 52 years old, respectively; *P*=0.00003; **Fig. 1f**) consistent with previous reports that these cryptic neoplasms are more common in older individuals[9-11]; and mutations in these samples showing highly similar VAFs to each other, consistent with their presence in the same clone (**Extended Data Fig. 2**). We removed 38 samples with evidence of neoplastic clones in the blood from further analyses (Methods).

After application of all filters, we obtained 163 mosaic mutations from 241 individuals, the large majority of which are likely to have arisen during early human embryogenesis (**Fig. 1g**; **Supplementary Table 2; Extended Data Fig. 3**). To confirm that such mutations are genuinely mosaic, we sequenced multiple single white blood cells from one individual and showed that the mutation analysed was only present in a subset **(Fig. 1h)**.

Most mutations of early embryonic origin would be expected to be present in all normal tissues and not just in white blood cells. From 13 individuals in whom putative early embryonic mutations had been detected in blood, we examined normal breast (composed of cells of ectodermal and mesodermal origins) and lymph nodes (composed of cells of mesodermal origin) for the presence of the early embryonic mutations. Most of the mutations were found in the additional normal tissues examined, with VAFs indicative of being mosaic and correlating with those found in blood, thus supporting the early embryonic origin of the mutations and further reducing the likelihood that they are due to neoplastic clonal haematopoiesis (Fig. 1i). The VAFs of the early embryonic mutations were lower in normal

breast and lymph node than in blood (Fig. 1i). This may be because, even at such an early stage of embryogenesis, there already exist different potentials of individual ICM founder cells to contribute to adult blood and other tissues but we cannot exclude other explanations (**Supplementary Discussion 1**).

In contrast to normal tissues, which are composed of multiple somatic cell clones, a breast cancer derives from a single somatic cell. Thus an early embryonic mutation would be expected either to be present in all cells of a breast cancer or in none, rather than in a proportion of cells as observed in blood and other normal tissues (although in practice the presence of contaminating non-cancer cells in the cancer sample has to be corrected for; Method). This was the pattern observed, with 37 mosaic mutations shared between the blood and the breast cancer from the same individuals, 105 non-shared and 21 uncertain, either due to a large deletion in the relevant region of the cancer genome (n=14) or statistical ambiguity (n=7) (**Figs. 2a, 2b**). The proportion of early embryonic mutations shared between the blood and the cancer is predicted to change according to the stage of early embryonic development at which the mutation occurred, with mutations acquired later being shared less often (**Extended Data Fig. 4**). Consistent with this expectation, embryonic mutations with lower VAFs in blood were shared less frequently with breast cancers (**Fig. 2c**).

These patterns of sharing of low VAF mutations in blood (which is of mesodermal origin) with normal and neoplastic breast tissue (which is of ectodermal origin) supports a model in which the most recent common ancestor (MRCA) cell of adult blood cells is the fertilized egg (**Fig 2d; Supplementary Discussion 2**), or is the MRCA cell of all/most somatic cells, rather than an alternative model of a single MRCA of the blood occurring at a later stage of embryogenesis with very restricted subsequent fate.

The VAFs of the 163 validated early embryonic mutations in blood, which ranged from 45% to 1%, provide insights into the early cellular dynamics of embryogenesis (**Fig. 3a**). If, in the large majority of embryos, the first two daughter cells of the MRCA cell of blood contributed equally to adult blood cells (symmetric cell doubling), a narrow 25% VAF peak would be

expected for mutations acquired at this stage. However, this peak was not observed indicating that asymmetric contributions are common.

To explore the basis of this asymmetrical contribution systematically, we generated a series of models of cell genealogies in which different branches contributed unequally to adult blood. The asymmetry that best fitted the observed VAF distribution is an average, across embryos, ~2:1 contribution of the first two daughter cells (cells I-1 and I-2; **Fig. 3b, 3c**). Moreover, this ~2:1 asymmetric cell contribution extends to two cells at the second cell generation (cells II-1 and II-2; **Fig. 3b, 3c**) and possibly to the third cell generation (Methods). The model with unequal contributions was clearly superior to a null model of symmetric cell doublings (P=1x10^{-40}, likelihood ratio test, **Fig. 3a, 3b**). This frequent unequal contribution of the earliest human embryonic cells to adult somatic tissues is consistent with previous indications from studies of mouse development[5,12-15].

The biological mechanism(s) underlying these asymmetrical contributions are not well understood. One daughter cell and its progeny may contribute more than the other because they intrinsically have a lower death rate, a higher proliferation rate and/or a preference for contributing to embryonic compared to extra-embryonic tissues[14-16]. Indeed, studies in mice have shown that cells separated from 2-cell embryos have different intrinsic developmental potentials[16,17]. Alternatively, the stochastic consequences of a bottleneck in early embryo development could be the source of the asymmetry. In the early blastocyst stage human embryo, which is composed of 50-100 cells (blastomeres), only the minority of cells (<20) present in the inner cell mass (ICM) eventually contribute to adult somatic tissues[18]. Under a model in which a small number (<20) of ICM founder cells are selected at random from a blastocyst composed of many (>50) equivalent blastomeres and most of the founder cells contribute to the specific adult cell types, it is likely that the progeny of the first two cells will, in many embryos, be selected in unequal proportions, as recently observed in mouse[19]. Indeed, our simulations indicate that stochastic allocation of early human embryonic cells into the ICM results in levels of asymmetric contribution similar to those observed, without any intrinsic differences of early cells (**Fig. 3d; Extended Data Fig 5;** Methods). Assuming the

stochastic hypothesis is correct, our simulation estimates that the number of ICM founder cells is approximately 10. Further studies will be needed, however, to clarify the source of the observed asymmetry in the contribution of early cells of the embryo to adult tissues.

Using the asymmetric cell-doubling model, we reconstructed the numbers of base substitutions present in each early embryonic cell. Taking into account the sensitivity of our low VAF mutation detection (**Method; Extended Data Fig. 3e**), approximately 14 substitutions are present with VAF 10%-35% in the blood of a person and we estimated a rate of 2.8 substitution mutations per cell per cell-doubling (**Fig. 4a**; **Supplementary Discussion 3**; 95% confidence interval 2.4-3.3) especially for the cell doublings at the first and second cell generations. A similar rate is obtained under a simple model without asymmetric contributions (**Fig. 4a;** Methods). This mutation rate per cell-doubling may not, however, directly equate to a rate per cell division because early embryonic development may involve cell loss, perhaps due to fatal chromosomal aberrations[6], and thus a cell-doubling may entail more than a single cell division. If cell loss is common in the first few divisions of life, the mutation rate per cell per cell division will be lower than the estimated rate per cell per cell-doubling cited above.

We then validated the early embryonic mutation rate using whole-genome sequences of bloods from three large families[20] (**Fig. 4b**). We found seven substitution mutations in children that were not present in their parents and that had features described above of early embryonic mutations (**Extended Data Fig. 6)**. Of these, four were on paternally derived and three on maternally derived chromosomes. We obtained a similar early embryonic mutation rate of 2.8 per cell per cell doubling (95% Poisson confidence interval 1.1-5.8; **Fig. 4a**).

The mutational spectrum of early embryonic mutations was predominantly C>T (42.9%), T>C (25.1%) and C>A substitutions (16.6%) (substitutions are referred to by the pyrimidine of the mutated Watson-Crick base pair) and was similar to that of *de novo* germline mutations (**Extended Data Fig. 7a**). Somatic mutations are caused by a diverse array of mutational processes including exogenous and endogenous mutagenic exposures, DNA modification, DNA editing and DNA maintenance mechanisms[7]. Each mutational process imprints a distinct

signature of mutations onto genomes on which it has been operative. We have previously developed methods for extracting mutational signatures from human cancers[7] (**Fig. 4c**) and have identified >30 different mutational signatures (http://cancer.sanger.ac.uk/cosmic/signatures). A combination of just two of these signatures, signatures 1 and 5, optimally accounts for the spectrum of early embryonic mutations, contributing 28% and 72% mutations, respectively (Fig. 4c). Signature 1 is thought to be due to spontaneous deamination of 5-methyl cytosine to thymine which results in C>T transitions primarily at CpG dinucleotides (**Extended Data Fig. 7b**). The mutational process underlying signature 5 is currently unknown. These two mutational signatures are found in almost all human cancers and the numbers of mutations associated with them exhibit strong positive correlations with age, suggesting that their underlying processes are endogenous and operate in most normal somatic cells at relatively constant rates throughout life[21]. Furthermore, signatures 1 and 5 appear to be the dominant mutational signatures contributing to human *de novo* germline mutations[20,21]. The current study therefore extends their domains of activity to early embryogenesis. We do not exclude the possibility that more mutational signatures, contributing smaller numbers of mutations, are operating in early embryogenesis, but a larger set of mutations will be necessary to identify them.

A very small number of early post-zygotic mutations have been previously reported[22-24]. This study therefore provides the first large-scale identification of early embryonic mutations with accurate VAF information, and illustrates the use of such information in elucidating developmental processes. The study reveals an average ~2:1 asymmetry of early human embryonic cells in terms of their contributions to adult tissues, thus providing insights into the fates of cells at very early developmental stages. Our results have also allowed the first estimate of the mutation rate and the mutational signatures and processes underlying base substitutions in the early human embryo, which appear to be comparable to those in mouse embryogenesis[5] and in adult somatic human tissues[20,25,26]. The early human embryonic mutation rate estimated here implies that, using similar methods to those introduced in mice previously[5], reconstruction of cell lineage trees using somatic mutations should be possible in humans.

References

1 Samuels, M. E. & Friedman, J. M. Genetic mosaics and the germ line lineage. *Genes* **6**, 216-237, doi:10.3390/genes6020216 (2015).

2 Erickson, R. P. Recent advances in the study of somatic mosaicism and diseases other than cancer. *Current opinion in genetics & development* **26**, 73-78, doi:10.1016/j.gde.2014.06.001 (2014).

3 Laurie, C. C. *et al.* Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature genetics* **44**, 642-650, doi:10.1038/ng.2271 (2012).

4 Ruark, E. *et al.* Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* **493**, 406-410, doi:10.1038/nature11725 (2013).

5 Behjati, S. *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* **513**, 422-425, doi:10.1038/nature13448 (2014).

6 Vanneste, E. *et al.* Chromosome instability is common in human cleavage-stage embryos. *Nature medicine* **15**, 577-583, doi:10.1038/nm.1924 (2009).

7 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

8 Oron, E. & Ivanova, N. Cell fate regulation in early mammalian development. *Physical biology* **9**, 045002, doi:10.1088/1478-3975/9/4/045002 (2012).

9 Genovese, G. *et al.* Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *The New England journal of medicine* **371**, 2477-2487, doi:10.1056/NEJMoa1409405 (2014).

10 Jaiswal, S. *et al.* Age-related clonal hematopoiesis associated with adverse outcomes. *The New England journal of medicine* **371**, 2488-2498, doi:10.1056/NEJMoa1408617 (2014).

11 Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nature medicine* **20**, 1472-1478, doi:10.1038/nm.3733 (2014).

12 Bruce, A. W. & Zernicka-Goetz, M. Developmental control of the early mammalian embryo: competition among heterogeneous cells that biases cell fate. *Current opinion in genetics & development* **20**, 485-491, doi:10.1016/j.gde.2010.05.006 (2010).

13 Plusa, B. *et al.* The first cleavage of the mouse zygote predicts the blastocyst axis. *Nature* **434**, 391-395, doi:10.1038/nature03388 (2005).

14 Zernicka-Goetz, M., Morris, S. A. & Bruce, A. W. Making a firm decision: multifaceted regulation of cell fate in the early mouse embryo. *Nature reviews. Genetics* **10**, 467-477, doi:10.1038/nrg2564 (2009).

15 Plachta, N., Bollenbach, T., Pease, S., Fraser, S. E. & Pantazis, P. Oct4 kinetics predict cell lineage patterning in the early mammalian embryo. *Nature cell biology* **13**, 117-123, doi:10.1038/ncb2154 (2011).

16 Bedzhov, I., Graham, S. J., Leung, C. Y. & Zernicka-Goetz, M. Developmental plasticity, cell fate specification and morphogenesis in the early mouse embryo. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **369**, doi:10.1098/rstb.2013.0538 (2014).

17 Morris, S. A., Guo, Y. & Zernicka-Goetz, M. Developmental plasticity is bound by pluripotency and the Fgf and Wnt signaling pathways. *Cell reports* **2**, 756-765, doi:10.1016/j.celrep.2012.08.029 (2012).
18 Hardy, K., Handyside, A. H. & Winston, R. M. The human blastocyst: cell number, death and allocation during late preimplantation development in vitro. *Development* **107**, 597-604 (1989).
19 Strnad, P. *et al.* Inverted light-sheet microscope for imaging mouse pre-implantation development. *Nature methods*, doi:10.1038/nmeth.3690 (2015).
20 Rahbari, R. *et al.* Timing, rates and spectra of human germline mutation. *Nat Genet*, doi:10.1038/ng.3469 (2015).
21 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nature genetics*, doi:10.1038/ng.3441 (2015).
22 Acuna-Hidalgo, R. *et al.* Post-zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation. *American journal of human genetics*, doi:10.1016/j.ajhg.2015.05.008 (2015).
23 Huang, A. Y. *et al.* Postzygotic single-nucleotide mosaicisms in whole-genome sequences of clinically unremarkable individuals. *Cell research* **24**, 1311-1327, doi:10.1038/cr.2014.131 (2014).
24 Dal, G. M. *et al.* Early postzygotic mutations contribute to de novo variation in a healthy monozygotic twin pair. *Journal of medical genetics* **51**, 455-459, doi:10.1136/jmedgenet-2013-102197 (2014).
25 Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 961-968, doi:10.1073/pnas.0912629107 (2010).
26 Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483-1489, doi:10.1126/science.aab4082 (2015).
27 Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54, doi:10.1038/nature17676 (2016).
28 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
29 Stephens, P. J. *et al.* The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400-404, doi:10.1038/nature11017 (2012).
30 Ju, Y. S. *et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. *eLife* **3**, doi:10.7554/eLife.02935 (2014).
31 Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome research* **22**, 568-576, doi:10.1101/gr.129684.111 (2012).
32 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
33 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).

34 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).
35 Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29**, 24-26, doi:10.1038/nbt.1754 (2011).
36 Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: a next-generation genome browser. *Genome research* **19**, 1630-1638, doi:10.1101/gr.094607.109 (2009).
37 Holstege, H. *et al.* Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. *Genome research* **24**, 733-742, doi:10.1101/gr.162131.113 (2014).
38 Marikawa, Y. & Alarcon, V. B. Establishment of trophectoderm and inner cell mass lineages in the mouse embryo. *Molecular reproduction and development* **76**, 1019-1032, doi:10.1002/mrd.21057 (2009).

METHODS

Samples and sequencing data

For initial discovery of early embryonic mutations, we analyzed whole-genome sequencing data from 304 blood samples of breast cancer patients which were sequenced as normal controls for the ICGC (International Cancer Genome Consortium) breast cancer study[27]. Genomic DNA was extracted from bulk white-blood cells collected from fresh peripheral bloods. Matched breast cancer samples for all the individuals were also analysed in parallel. Of these, 25 samples with putative DNA contamination were removed (see below for more details), and 279 samples were used for the detection of early embryonic mutations (the sample information is available in Supplementary Table 1). For validating the mutation rates, we also used whole-genome sequencing data from 19 blood samples from 3 families[20]. For confirmation of early embryonic mutations in non-blood normal tissues, we extracted genomic DNA from FFPE (formalin-fixed and paraffin embedded) lymph nodes and normal breast tissue surgically resected during mastectomy procedures. The whole-genome sequencing data analysed in this study were generated using Illumina platforms (either Genome Analyzer or HiSeq 2000). Sequencing reads were aligned to human reference genome build 37 (GRCh37) using the BWA alignment tool[28]. All PCR duplicate reads were removed.

DNA contamination control

We thoroughly checked for possible sources of DNA contamination: tumour-normal swap; matched tumour DNA contamination in blood; and cross-contamination with DNA from other individuals. Cases of tumour-normal sample swap were identified by examining the presence of genome-wide copy number variations in the putative normal samples. Cases of matched tumour DNA contamination were identified by examining the VAFs in the blood sequencing data for the somatic substitution variants identified in the matched cancer using CaVEMan software[29] (available at https://github.com/cancerit/CaVEMan/). When the average VAF of cancer specific substitutions was more than 1% in a blood sample, we regarded the blood sample to be contaminated by a matched tumour DNA sample. Finally, for each sample, the

level of DNA cross-contamination with tissue from other individuals was estimated as described previously[30].

Variant calling

VarScan2 software[31] was used for initial early embryonic variant calling. Input vcf files were generated from whole-genome sequencing bam files using *samtools*[32] *mpileup* with three options -q 20, -Q 20 and -B. Then VarScan2 *somatic* was applied to blood samples with matched tumour samples as reference. Three options were applied for the VarScan2 running, --min-reads2 4, --min-ave-qual 20, and --strand-filter. We selected substitution variants with VAFs ranging from 0.1 to 0.35 as putative early embryonic mutations. We removed putative mutations near germline indels (within 5bp), because these are mostly false positives due to mismapping. Putative mutations likely to be sequencing artifacts and/or germline polymorphisms were removed if the variants were also present in the unmatched blood samples analysed in this study, or were known germline polymorphisms with at least 1% population allele frequency identified from the 1000 Genomes Project (Nov.2013), or deposited in dbSNP (v138). We removed putative variants in segmental duplications, simple repeats, repetitive sequences (RepeatMasker) and homopolymer sequences in the human reference genome (downloaded from UCSC genome browser, http://genome.ucsc.edu/).

Substitution phasing

We phased the putative embryonic variants to heterozygous germline substitutions using sequences from whole-genome sequencing as described previously[30,33]. For more conservative phasing, we did not use sequences at the 4bp extremes of each read, where substitutions and indels are not well called. From blood whole-genome sequencing data, we classified the putative variants into 4 groups, 'phasing not available', 'mixed pattern', 'no evidence of subclonality' and 'subclonal' using criteria as follows:

(1) Phasing not available: no available read covering both the mutation and the heterozygous SNP in the vicinity

(2) Mixed pattern: the putative variant is present in both the bi-allelic haplotypes of heterozygote SNPs

(3) No evidence of subclonality: the putative variant is completely and exclusively present on one of the two haplotypes of heterozygote SNPs

(4) Subclonal: the putative variant is present in a fraction of one of the two haplotypes of heterozygote SNPs. The variant is not present on the other haplotype.

Putative mutations categorized other than subclonal were removed. For the subclonal mutations, we estimated the probability of false subclonality due to sequencing errors. For this calculation, we counted only informative reads, which were participating in the phasing: reads covering the putative mutation locus and one of the alleles of the inherited heterozygous SNP in which the early mutation is linked.

$$P_{error} = \prod_{i}^{V}(Q1_i + Q2_i - Q1_i \cdot Q2_i) + \prod_{j}^{W}(Q1_j + Q2_j - Q1_j \cdot Q2_j) - \prod_{i}^{V}(Q1_i + Q2_i - Q1_i \cdot Q2_i) \times \prod_{j}^{W}(Q1_j + Q2_j - Q1_j \cdot Q2_j)$$

Q1 and Q2 are sequencing error rates of the bases at the putative mutation and the heterozygote SNP loci, respectively; i represents each of the informative reads harboring the mutant base at the early embryonic mutation site; V is the total number of informative reads with the mutant base; likewise, j represents each of the informative reads harboring a wild-type base at the early embryonic mutation site and W is the total number of such reads. When there was more than one heterozygous SNP site that was used for phasing, we calculated a string of phasing error rates (P_{error}) from every SNP site and multiplied them to obtain an overall phasing error rate.

Copy number of mutation loci

We removed any putative mutation if it was located in a region with copy number higher than two. We isolated potential copy number variation of each genome using both intra-sample and inter-sample methods. For the intra-sample method, we calculated the standard deviation of read-depth from all (~2 million) germline heterozygous SNP sites from every normal whole-genome sequencing dataset. When the local coverage of an early embryonic mutation candidate was higher than the 95% percentile (i.e. local depth is greater than genome-wide mean WGS coverage + 1.645 x stdev; it is ~46x in typical 30x coverage sequencing) of the

sample, we considered the site was possibly duplicated thus removed from our further analyses (**Extended Data Fig. 8a**). For the inter-sample method, we clustered the normalized normal WGS read counts of a candidate region (from 1kb upstream of the mutation site to 1kb downstream) from all the samples included in this study. If the normalized copy number of the target sample was either an outlier in the clustering or was two times higher than expected from genome-wide average, the mutation candidate was considered to locate in a germline copy number variant region and thus filtered out (**Extended Data Fig. 8b, 8c**).

Mutations shared by the paired tumour tissue

Then we investigated whether the early embryonic mutation candidates were also present in cells of the breast cancer from the same individual. This is not always straightforward because (1) whole-genome sequencing of cancer tissue generates a mixture of sequences from cancer and contaminating normal cells and (2) copy number changes are quite frequent in the cancer genome. Using the ASCAT algorithm[34], based on analysis of the variant allele fraction for heterozygous germline SNPs for regions departing from diploidy in the tumour genome, we estimated the tumour cell fraction ('f' in the formula below), ploidy of cancer genome ('p') and local A (major) and B (minor) allele copy numbers ('a' and 'b', respectively). Each mutant allele was previously phased to either A or B allele nearby. Using these estimates, we built a model for the expected number of reads (N) supporting the mutant allele in paired-cancer genome sequencing in three different scenarios:

I) The mutant allele is not shared (and approximate 95% binomial confidence interval),

$$N = D\pi_0 \, , (95\%\ CI: 1.96\sqrt{D\pi_0(1-\pi_0)}\ )$$

, D is the read-depth of the mutant site in matched cancer WGS sequencing and

$$\pi_0 = {2(1-f)\rho}\Big/{\big((a+b)f + 2(1-f)\big)}$$

, ρ is the expected VAF of the mutant allele.

II) The mutant allele is phased to B allele (with 95% confidence interval),

$$N = \ D\pi_1, (95\%\ CI: 1.96\sqrt{D\pi_1(1-\pi_1)})$$

, $\pi_1 = {(fb + 2(1-f)\rho)}/{\big((a+b)f + 2(1-f)\big)}$

If $n_B = 0$ we cannot differentiate scenario I and II (loss-of-mutant allele).

III) The mutant allele is phased to A allele (with 95% confidence interval),

$$N = D\pi_2, (95\%\ CI: 1.96\sqrt{D\pi_2(1-\pi_2)})$$

, $\pi_2 = {(fa + 2(1-f)\rho)}/{\big((a+b)f + 2(1-f)\big)}$

According to these models, we assigned our mutation to four groups: 'non-shared' (model I), 'shared' (model II or III), 'loss-of-mutant allele' (when the mutant allele is phased to B allele and b is 0) and 'uncertain' (when more than 1 model could explain or no convincing ASCAT result is available for the sample).

Visual inspection

We visually inspected all the candidate embryonic mutations using the Integrative Genomic Viewer[35] and JBrowse[36]. We confirmed that genomic regions with putative embryonic mutations were not in sequences with evidence of artifacts and thus that any putative mutation was supported by high quality sequencing reads. Two examples of early embryonic mutations are shown in Figs. 2a and 2b.

Validation by MiSeq amplicon sequencing

We tried to validate all the putative early embryonic mutation sites. We designed 959 pairs of PCR primers (**Supplementary Table 3**) for 863 candidate early mutations to make amplicons for the putative mutation sites along with the nearby heterozygote SNPs used for phasing from the blood and paired-cancer DNA samples of the individual harboring the putative mutation. After clean-up using ExoSAP-IT (Affymetrix Inc., Santa Clara, CA, USA), all amplicons from blood and matched cancer tissues were separately pooled and sequenced by 2 x 250bp MiSeq sequencing (Illumina Inc., San Diego, CA, USA) 2 runs per pool, expecting > 1000x coverage per amplicon. Because the read-depth is very high in amplicon sequencing, we could obtain a much more precise variant allele fraction of the putative embryonic mutation along with accurate phasing to the germline heterozygote substitution. The VAFs for germline heterozygote substitutions in non-repetitive genome regions showed a

clear peak at 0.5 (**Extended Data Fig. 9a**). To estimate the extent to which the amplification process biased the VAFs, we fitted a beta-binomial distribution with mean 0.5 and dispersion to the numbers of reads supporting both alleles in heterozygous SNPs (which have an expected VAF=0.5). This confirmed that the additional uncertainty introduced by amplifications was very small (θ = 223.88, overdispersion $\rho = 1/(1+\theta)$ = 0.004). This estimate of the overdispersion was used in the maximum likelihood asymmetric models. The targeted amplicon sequencing showed high precision in the assessment of the VAF of a mutation (**Extended Data Fig. 9b)**. The MiSeq validation experiment confirmed that the candidate mutations were not sequencing artifacts nor inherited mutations both from the resulting VAFs (ranged from 0.01 to < 0.5, mostly < 0.35) and from phasing to the local heterozygous SNP. From this validation study, we found that there is a clear linear relationship between phasing error rates (as calculated above) and validation success rate (**Extended Data Fig. 10**). We could not create amplicons from some mutation candidates due to lack of DNA samples or unsuccessful PCR reactions. Finally, we rescued 14 early embryonic mutations because they are likely to be true on the basis of phasing error probability (**Supplementary Table 2**).

Mutation validation using single cells

From the blood of one individual (PD7344) we sorted 144 granulocytes. Genomic DNA of each single cell was extracted and whole-genome amplified using the REPLI-g Single Cell Kit (Qiagen Inc.) using the manufacturer's protocol. Of the 144 single cells, 131 provided substantial amounts of WGA DNA. PCR amplicons were produced targeting the early embryonic substitutions in the sample (chr3:187268541 C>A). PCR reactions were successful from 118 WGA DNAs. After clean-up of the 118 PCR products, capillary sequencing was performed. Of these, 41 showed allelic dropout of the DNA haplotype on which the embryonic mutation was present (absence of the T allele of rs17726238) and thus were not further considered. Among the 77 informative amplicon sequencing results, 24 showed clear evidence of the embryonic substitutions as shown Fig 1h.

Late somatic mutations due to clonal haematopoiesis

Age-related clonal haematopoiesis is quite common, and observed in more than 10% of persons older than 65 years old[9-11]. Like mutations that have occurred in the very early embryo, these late mutations appear to be subclonal (mosaic) in adult blood. However, such late mutations are rarely shared with the breast cancer sample from the same individual because the vast majority of them occurred after formation of the three germ layers, specifically in the mesodermal lineage. In addition, late clonal expansions in the blood invariably carry a large number of co-clonal mutations accumulated throughout life[37], and so many subclonal mutations with similar VAFs are detected together in the blood sample. In this study, we found that each blood sample harbors a median of 1 validated phased subclonal mutation. According to their distribution (**Fig. 1c**), we regarded 31 samples with at least 5 validated subclonal mutations as outlier samples, defined as deviating from the median value by more than twice the interquartile range. Consistent with the hypothetical presence of late clonal expansions in these outlier samples, the proportion of non-shared mutations abruptly increases from this point (**Fig. 1c**). Furthermore, we searched 73 cancer genes which have been reported to drive clonal haematopoiesis[9-11] for low VAF somatic mutations (supported by at least 3 mismatches) and identified eight samples with mutations in *DNMT3A*, *ASXL1*, *JAK2*, *PTPN11* and *CBL* genes. Of these, four samples were found among the 31 outlier samples. Conservatively, the remaining four samples were also classified as containing clonal haematopoiesis despite the small number of mutations found in them, and therefore removed from downstream analyses. Finally, we assessed whether mutation candidates obtained from each sample showed significantly similar VAFs to each other compared to the other samples, indicating that those mutations may be present in same blood clone, and thus filtered out three additional samples. Indeed, from the 38 filtered samples, we observe that mutations have more similar VAF to the other mutations in the same sample (calculated by $VAFi/\overline{VAF}$, where I represents each mutation in the sample) compared to the mutations in samples with 2-4 mutations (**Extended Data Fig. 2**). As a result, out of the total 279 samples, we classify 241 samples as having no evidence of clonal haematopoiesis, and therefore informative for detecting embryonic mutations.

Finally, we assessed whether matched tumour sequences showed evidence of the mutant allele with significantly higher VAFs than background sequencing error rate levels (**Extended**

Data Fig. 9c). This would be expected, because normal cells are always present in cancer samples and a fraction of the normal cells would carry the mutant allele if a mutation is truly embryonic origin. Fifteen candidate mutations, which were not found in the matched tumour samples, were removed through this step. After application of all filters, we identified 163 likely early embryonic mutations from 241 samples.

Asymmetry in early cell doublings

In order to fit different lineage models to the VAF of embryonic mutations, we used a likelihood approach. If read counts were fully independent, allelic counts from each mutation could be modelled as being binomially distributed. However, to account for the overdispersion caused by the amplification process prior to library preparation, we assume allelic counts to be beta-binomially distributed. As shown above, we estimated the overdispersion parameter θ=223.9 ($CI_{95\%}$: 201-248). Over 98.7% of heterozygous SNPs had a VAF in the range [0.4,0.6] in the re-sequencing dataset (**Extended Data Fig. 9a**)

If the first cell doubling gives rise to two daughter cells that contribute equal numbers of cells to the adult (or the adult blood population), the doubling is considered symmetrical. Otherwise, the doubling is considered asymmetrical, with one cell contributing a fraction α_1 of the cells in the adult and the other cell $1-\alpha_1$. Assuming that embryonic mutations are heterozygous, the expected VAF of a mutation occurring in branch 1 of the lineage is $0.5*\alpha_1$ and in branch 2 is $0.5*(1-\alpha_1)$. The same applies to any doubling in the lineage, with the two daughter cells contributing α_n and $1-\alpha_n$, relative to the contribution of the mother cell (n). This allows calculating the expected VAFs in the adult cell population for mutations occurring at each branch of the model lineage tree (v_b).

For each embryonic mutation, j, we observe the number of mutant reads (m_j) and the total coverage at the site (c_j). The likelihood of observing a given mutation under a particular lineage model requires integrating the likelihood of observing the mutation under each branch of the lineage, considering also the mutation rate at each branch and the sensitivity to mutations from each branch. In other words, the VAFs are fitted to a mixture model as

mutations could have occurred at any branch in the tree. The total log-likelihood of the model is the sum of the log-likelihoods from all mutations.

$$\prod_{j=1}^{N} \frac{1}{\sum_{b=1} r_b * s_b} \sum_{b=1}^{B} BetaBin(m_j, c_j, v_b, \theta) * r_b * s_b$$

Where *N* is the total number of mutations in the dataset (*N*=163), *B* is the total number of branches in the model and r_b is the (relative) mutation rate of the branch. s_b is the (relative) sensitivity to mutations from the branch, which is a function of the expected VAF of mutations from the branch (v_b). Sensitivity as a function of VAF is calculated as described in the section below.

Statistical comparison of models of increasing complexity

In order to evaluate whether a lineage with one asymmetric doubling fits the data significantly better than a symmetric model, we obtained the maximum likelihood estimate for α_n from each of the 15 doublings from the first 4 cell-generations while keeping all other doublings symmetrical. The best 1-asymmetric-rate model is tested against the symmetric model with a likelihood ratio test with 1 degree of freedom, and the p-value is subjected to Bonferroni multiple testing correction to account for the 15 models evaluated. This revealed that a lineage where the first doubling is asymmetric with $\alpha_1 \approx 0.61$ fits the data much better than a symmetric model (LL0=-1444.4, LL1=-1366.3, $P < 10^{-16}$).

In order to test models with additional asymmetric rates we used a heuristic approach. Instead of testing all possible combinations of asymmetric rates, we tested the impact of adding an extra asymmetric rate to the previous model (14 alternative models). The best model included asymmetry in the cell doubling of the dominant daughter cell in the first cell doubling (LL1=-1366.3, LL2=-1349.102, Bonferroni-corrected *P*=3.1e-08). The same approach was used to find a better model with three and four asymmetric doublings. The best model with three asymmetric doublings is only marginally better than the best model with two

asymmetric doublings (LL3=-1344.784, Bonferroni-corrected *P*=0.021). More complex models provided no significantly improved fits to the data.

In order to evaluate whether other asymmetric lineages with two or three asymmetric rates could provide better fits, we exhaustively calculated the maximum-likelihood values of all possible lineages with two or three asymmetric doublings in the first four cell-generations. No model provided a better fit to the ones found by the heuristic approach. This analysis strongly supports a lineage with at least two asymmetric rates (first and second branches).

The confidence intervals shown in Fig 3c were calculated by non-parametric bootstrapping (*i.e.* resampling the original data with replacement) followed by numerical search of the maximum likelihood values of the top seven rates in the lineage.

Estimating the average mutation rate from asymmetric lineage models

Assuming a given lineage model, a global estimate for the average mutation rate per genome per doubling in the early embryo can be obtained with the following equation:

$$\frac{N}{S\sum_{b=1} s_b}$$

N is the total number of embryonic mutations detected (*N*=163), *S* is the number of samples studied (*S*=241) and s_b is the sensitivity to detect a mutation from a particular branch of the lineage tree. Further, an approximate estimate of the average mutation rate at different cell generations could be obtained using an Expectation-Maximisation (EM) algorithm. These estimates may be more robust against possible contamination from neoplastic expansions at very low VAFs than the global estimate above.

Assuming a particular lineage, the relative probability (*expectation step*) of a mutation (j) coming from one particular branch (b) is given by:

$$p_{b,j} = \frac{BetaBin(m_j, c_j, v_b, \theta) * r_b * s_b}{\sum_{i=1}^{B} BetaBin(m_j, c_j, v_i, \theta) * r_i * s_i}$$

In the first iteration of the EM algorithm, the mutation rate (r_j) of all branches is considered identical. The number of mutations estimated to come from each branch is then calculated as the sum of these probabilities across all mutations:

$$N_b = \sum_{j=1}^{N} p_{b,j}$$

N_b is then used to update the mutation rate per branch (*maximisation step*). And these two steps are iterated until convergence, obtaining an improved fit to the data and estimates of the mutation rates per branch. To constrain the parameters of the model, the rates of all branches from the same cell-generation are maintained identical during the EM algorithm. Confidence intervals were obtained by bootstrapping (400 replicates). Importantly, allowing the mutation rates of the first three cell-generations to vary freely with respect to the rest of the lineage (values shown in main text, Fig. 4a), does not significantly improve the fit of the model (LL=-1347.0 as opposed to LL2, p-val=0.24, 3 degrees of freedom).

Simulation of sensitivity

We estimated the sensitivity for early embryonic mutations from simulation studies. The sensitivity will be dependent on the target VAF (ρ) of early mutations. First, we randomly generated 1,000 *in-silico* embryonic mutations genome-wide. *In-silico* mutations within known gaps of the human reference genome were removed and replaced by newly generated mutations. Note that this means that sensitivity and so the mutation rates estimated in our study exclude mutations present in gaps, which approximately correspond to ten percent of the human genome. Next, under 21 different theoretical VAF (ρ; 0.016, 0.028, 0.031, 0.056, 0.063, 0.083, 0.111, 0.125, 0.139, 0.167, 0.194, 0.222, 0.250, 0.278, 0.306, 0.333, 0.361,

0.389, 0.417, 0.444, 0.472) we queried how many of them could be detected on average from the whole-genome sequences of 241 samples. The same filtration steps for real mutation candidates were applied for the *in-silico* mutations: if mutations are found in 1000 Genomes Project dataset, dbSNP variation, segmental duplications, simple repeats, repetitive sequences by RepeatMasker, homopolymers, and potential copy number gain regions, we regarded these mutations as undetectable. Then, for each potentially detectable *in-silico* mutation, and under several given ρ, we calculated the fraction of mutations that could be successfully detected and successfully phased to at least one heterozygous SNP nearby in each individual WGS.

$$P(observed|\rho) = P(detection|\rho) \cdot P(phasing|\rho)$$

where P(detection|ρ) is the probability of a mutation having a sufficient number of reads supporting the mutant allele (at least 4, or the cutoff value in this study) and a VAF within the range considered in the discovery phase of this study (from 10% to 35%). Likewise, P(phasing|ρ) represents the probability of successful phasing a mutation to the heterozygous SNP nearby. We calculated P(detection|ρ) and P(phasing|ρ) as below:

$$P(detction|\rho) = \sum_{r=\max(4,roundup(0.1D))}^{roundoff(0.35D)} \binom{D}{r} \rho^r (1-\rho)^{(D-r)}$$

$$P(phasing|\rho) = 1 - \prod_{i}^{\max(1,N)} ((0.5+\rho)^{Si} + (1-\rho)^{Si} - 0.5^{Si})$$

where roundup () and roundoff() functions round to the higher or the closest integer number, respectively. D is the read-depth of each detectable *in-silico* mutation site, N represents the total number of heterozygous SNPs which are available for phasing, i is each of the heterozygote SNPs and Si is number of reads spanning both a mutation locus and the heterozygous SNP. For simplicity of simulation, we assumed all the bases have a good base quality (i.e. phred score >20). Finally, we added all probabilities, P(observed|ρ), obtained from an individual given ρ. When ρ is fixed, P(observed|ρ) correlates with read-depth of blood whole-genome sequencing, and the regression line was obtained using loess regression. We obtained our sensitivity estimates for the 21 different ρ values using this approach and a simulated coverage of 32-fold coverage (median coverage for 241 blood samples). For

example, 4.41% of the 1000 *in-silico* mutations with ρ=0.25 were detectable when whole-genome sequencing coverage was 32x (**Extended Data Fig. 3e**).

A stochastic model of embryoblast formation

In the maximum likelihood fitting of lineage models described above, a single lineage tree was fitted to the data from multiple different individuals. The resulting lineage intends to be a merely descriptive representation of the average contribution of different cells across embryos. The model implicitly assumes that the same asymmetric lineage describes all patients and that the first divisions of the embryo follow a largely constant pattern across individuals. It remains unclear whether early embryonic development in viable embryos under physiological conditions follows a strict plan in humans or whether there is extensive variation between individuals, as observed in mouse[19]. In the presence of extensive variation in the early lineage across embryos, the asymmetry rates estimated using a constant lineage should be interpreted with caution.

Interestingly, extensive asymmetry in the contribution of the first cells of the embryo to the adult cell pool can also emerge under more stochastic models of embryo development. As a proof-of-principle, here we show how a bottleneck in the pre-implantation embryo, in which only a randomly selected subset of cells contributes to the final somatic tissues, can give rise to extensive asymmetry in the contribution of the first few cells of the embryo to the adult cell pool, not dissimilar to the general patterns observed in this study.

All final embryonic tissues are thought to derive from a fraction of cells in the blastocyst termed the inner cell mass (ICM), while the rest of the blastocyst (the trophoblast) will form the placenta and other extra-embryonic supporting tissues, and will not contribute to the adult cell pool. In mice this separation is thought to involve about 12 ICM cells gravitating at the center of the blastocyst at the 32-cell stage[38]. This imposes a significant bottleneck to the contribution of the first few cells in the embryo to the adult cell pool. Let us consider a simple bottleneck model where a completely random subset of *l* cells from the *n*-cell stage embryo are selected to form the adult cell pool. If there were m cells carrying an early somatic

mutation out of a total of n, the probability to subselecting k in a draw of l cells is given by the hypergeometric distribution. This is to be multiplied by the probability that m cells are mutated due to early germline mutations. Without a bottleneck, variant alleles would only be expected at powers of ½, with intensities following an 1/f power law due to the increase in the number of cells with every cell doubling. Hence the probability of selecting k mutated cells out of a total n cells is given by:

$$P(k;l,n) = \Sigma_{0<= i <= 2^n} \text{phyperg}(k;\, l,\, m = 2^n\text{-}2^i,\, n=2^i) \times 2^i / \text{const} \quad (1)$$

where const is a normalisation constant. Note that this distribution has support on VAF k/l, rather than $1/2^i$. The latter is approached in the limit that $l = n$, that is that all cells would propagate to the final somatic tissue (**Extended Data Fig. 5a**). The overall probability of observing mutations at a given VAF v is then to be multiplied by the sensitivity $S(v)$ to detect mutation a given frequency, and the additional dispersion arising from detecting mutations on a finite number of x sequencing reads at a given coverage c, modeled by a beta-binomial sampling model, as described in the deterministic modeling used in the previous sections.

$$p(x;l,n,c) = \Sigma_k P(k;l,n) \times S(k/l) \times \text{pbetabin}(x;\, \text{prob} = k/l;\, \text{disp} = \rho) / \text{const} \quad (2)$$

, the dispersion ρ is inferred from heterozygous SNPs and taken to be θ=223.9, ρ = 1/(1+ θ).

We may hence fit the likelihood (2) to the observed data, knowing the number of mutated reads x and coverage c for each patient, given the number of ICM cells l and cells n. The maximum likelihood is obtained for l=11 ICM cells separating after 6 generations, or n=64 cells (**Extended Data Fig. 5b**), although there are many solutions with similar likelihood.

From Eq. (2), an estimate of the overall histogram p(v) can be computed as the average over all data points $p(v;\, l,\, n) = \Sigma_i\, p(x_i = vc_i;\, l,\, n,\, c_i) \,/\, N$, where N = 163 is the number of observations. Using a Bayesian approach, assuming a uniform prior on the number of cell generations at which ICM commitment occurs ranging from 3 to 8, and similarly a uniform

prior on the number of ICM cells ranging from 5:32, allows for computing the posterior probability of the observed data as:

$$p(v) = \Sigma_l \Sigma_n p(v;l,n) \times p(l) \times p(n) \quad (3)$$

The result is shown in **Extended Data Fig. 5c**.

This model shows how a simple random selection of a subset of the cells in the early embryo can lead to substantial asymmetries in the contribution of the first few cells in the embryo to the final adult cell pool. We note that this represents one extreme of possible combined deterministic and stochastic scenarios. It remains unclear to what extent viable embryos under physiological conditions follow a tightly predetermined developmental plan or whether largely stochastic processes dominate before the formation of the first structures in the blastocyst. The available data cannot distinguish between these models, but we anticipate that more detailed analyses of early embryonic somatic mutations could shed some light on this question. In particular, deterministic models predict that all individuals will share a very similar lineage pattern while stochastic models predict largely different early lineages among individuals.

Family analyses

Genomic DNA was extracted from peripheral blood of 19 individuals from three large families. We detected subclonal substitutions in 13 children using identical methods for the blood tissues of 241 breast cancer patients, i.e. DNA contamination control, variant calling, phasing to nearby heterozygous SNP, assessment of copy number of the mutation loci, and visual inspection as described above. We detected 7 early embryonic mutations (Extended Data Fig. 6), which were subclonal and not shared by the parents or any siblings, therefore these are highly likely to be post-zygotic mutations which occurred at the early embryonic stages of a specific child.

We calculated the rate of early mutations from families (R_{family}) as below:

$$R_{family}/_{R} = \frac{\frac{N_{family}}{S_{family}} \cdot \alpha}{\frac{N}{S}}$$

Where R is the overall average early mutation rate (2.8 mutations per cell per cell generation), N is the number of mutations (n=163) and S is the total sample size (n=241). Likewise, N_{family} is the number of mutations (n=7) identified from family data and S_{family} is the total number of children analysed (n=13). α is relative sensitivity of early mutations in family data, which must be less than 1 because sequencing coverage is ~7x coverage lower in families (25x) than the unrelated 241 blood samples (32x). The simulation of sensitivity (shown above) suggests that α is 0.796. A Poisson Exant test was used to calculate the 95% confidence interval of R_{family}.

Detecting contributions of mutational signatures

Mutational signatures were detected by refitting of previously identified and validated consensus signatures of mutational processes (http://cancer.sanger.ac.uk/cosmic/signatures). All possible combinations of at least seven mutational signatures were evaluated by minimizing the constrained linear function:

$$\min_{Exposures_i \geq 0} ||\overrightarrow{DeNovoMutations} - \sum_{i=1}^{N} (\overrightarrow{Signature_i} * Exposure_i)||$$

Here, $\overrightarrow{DeNovoMutations}$ and $\overrightarrow{Signature_i}$ represent vectors with 96 components corresponding to the six types of single nucleotide variants and their immediate sequencing context and $Exposure_i$ is a nonnegative scalar reflecting the number of mutations contributed by this signature. N reflects the number of signatures being re-fitted and all possible combinations of consensus mutational signatures for N between 1 and 7 were examined, resulting in 2,804,011 solutions. Model selection framework based on Akaike information criterion was applied to these solutions to select the optimal decomposition of mutational signatures. The analysis revealed that signature 1 and signature 5 best describe the set of embryonic mutations. Including any other mutational signature did not improve the explanation of the set of embryonic mutations.

Figure Legends

Figure 1. Detection of somatic mutations acquired in early human embryos.

(a) Phylogenetic tree of early embryonic cells and transmission of an early embryonic mutation. Each white-filled circle represents an embryonic cell. The diploid genome is represented by two black bars inside the white circles. An early embryonic mutation occurring in a cell at the 2-cell stage is represented by a red square.

(b) Early embryonic mutations appear as somatic mosaicism in normal polyclonal tissue, for example, blood. The mutations are found on a subset of sequencing reads from bulk tissue.

(c) Distribution of the numbers of early embryonic mutations per individual genome showing all samples (red bars), and the finally selected samples with no evidence of haematopoietic clonal expansions (blue bars). The proportion of non-shared mutations (not present in breast cancers) is represented by the green line. Error bars denote 95% confidence intervals (binomial test).

(d-e) Early embryonic mutations are either absent from cancer cells ('non-shared'; d) or are fully clonal mutations in the cancer cells ('shared'; e) depending on their embryonic cellular origins.

(f) The median age of individuals with evidence of neoplastic expansion in blood is 12 years higher than individuals without it.

(g) 163 early embryonic mutations identified from 241 individuals are represented by dots. The 23 human chromosomes are shown in the outer layer (excluding chromosome Y). The vertical axis in the inner layer represents the VAF of the mutation. Sharing with cancer cells and the substitution types are represented by the shapes and colours of the dots.

(h) Single cell genome sequencing shows that a fraction of blood cells harbor a mosaic mutation (PD7344b, chr3:187,268,541 C>A).

(i) Early mutations are also found in non-blood normal tissues with VAFs indicative of being mosaic and correlating with those found in blood.

Figure 2. Early embryonic mutations identified from sequencing of adult blood tissues paired with matched tumour tissue sequencing.

(a) An example of an early embryonic mutation not shared with cancer cells. In the blood of patient PD6416, a G>A mutation at chr8:5,161,217 (highlighted in yellow) is present with a VAF of ~14% from both whole-genome and high coverage targeted-amplicon sequencing. This suggests that ~28% of adult white blood cells harbour the mutation. Consistent with its somatic origin, the mutation is present on a subset of reads that include variant alleles of two nearby germline SNPs (rs7009390 and rs7009505). This mutation was absent in the breast cancer cells from the same patient. The minimal low VAF (2.6%) observed in the targeted amplicon sequencing of the tumour is consistent with normal cells contaminating the tumour sample.

(b) An example of an early embryonic mutation shared with cancer cells. In the blood of patient PD7211, a G>A mutation at chr6:161,658,557 (highlighted in yellow) is present with a VAF of ~20% in blood suggesting that ~40% of blood cells have the mutation in their genomes. The mutation is present in the cancer sample with a VAF of 42.1% in targeted amplicon sequencing, consistent with its presence in all cancer cells and a contaminating population of non-neoplastic cells that lack the mutation.

(c) The proportion of mutations shared between the blood and the cancer correlates with the VAF of mutations in blood (from 1st quartile (low VAF, later mutations) to 4th quartile (high VAF, earlier mutations)), consistent with expectations for early embryonic mutations.

(d) The observed proportion of shared mutations (26%; red bar) indicates that the MRCA cell of blood is likely to be the same as the MRCA cell of many or all somatic tissues and/or the fertilized egg. The four blue bars show the expected proportions of mutations shared between the cancer and blood when there are 0, 1, 2, or 3 cell generation gaps respectively between the MRCA of the blood and the MRCA of the breast cancer (Extended Data Fig. 3). Binomial tests, * $P < 0.01$; ** $P < 10^{-9}$; *** $P < 10^{-15}$.

Figure 3. Asymmetric average contributions of early embryonic cells to adult somatic tissues.

(a) The distribution of VAFs of 163 early embryonic mutations in blood. Light green bars, VAF estimated from targeted ultra-high depth amplicon sequencing; gray bars, VAF estimated from whole-genome sequencing when targeted amplicon sequencing was not possible. The expected VAF distributions (with adjustment for sensitivity of mutation detection) from symmetric (equal contributions) and best-fitting asymmetric cell doubling model (unequal contributions) are shown with black and red lines, respectively.

(b) A contour plot showing the optimization of asymmetries in cell doublings. The horizontal axis and vertical axis present the asymmetry levels for the first and the second dominant cell doublings (cell doubling of MRCA and I-1 cells, respectively). Compared to the symmetric model (black arrow), the maximum likelihood asymmetric model (red arrow) provides a much better fit to the data ($P=1x10^{-40}$, Likelihood Ratio Test).

(c) Maximum likelihood estimates of the average contribution of different cells to the adult cell pool. The genealogy of early ancestral cells of adult blood cells is shown. The final contributions of each early cell to adult tissues are presented as pie graphs in which the red slices indicate the contributions to adult blood of each embryonic cell. These estimated asymmetric rates are population averages and could emerge through deterministic developmental plans or through stochastic processes. The asymmetries of each cell doubling are shown in blue (significantly asymmetric) or gray (the possibility of symmetric doubling is not excluded).

(d) The relative contributions of the first four cells (cell II-1 (blue), II-2 (light blue), II-3 (orange) and II-4 (red)) to adult somatic cells under a stochastic bottleneck model, in which a small number of cells of the early embryo (x-axis) are randomly selected to form the ICM (see Methods).

Figure 4. Rates and mutational signatures of somatic mutations in the early human embryo.

(a) Estimates of early embryonic mutation rates. Mutation rates from 163 early embryonic mutations calculated from the best-fitting asymmetric model (top green panel). The overall

rate and rate estimates for each cell generation are shown with red and black dots, respectively. Broken lines represent 95% confidence intervals obtained by bootstrapping (see Methods). Equivalent mutation rates from the symmetric model (middle orange panel). Mutation rate estimates from 3 families (bottom blue panel).

(b) Early embryonic mutations from whole-genome sequencing of 3 large families. Pedigrees are shown. Seven *de novo*, low VAF mosaic mutations were found in 13 children. Each mutation is shown with a number (index) inside the white rectangles or circles in the pedigrees. The maternal/paternal origins of the chromosomes of the mutations are shown in red (maternal) or blue (paternal). One of the mutations (#5) in family 569 is shown in more detail with an IGV image (highlighted in blue). Daughter 4 has a *de novo* subclonal C>T mutation (chr9:120,446,887) on the paternally transmitted chromosome 9.

(c) The mutational spectrum for 163 early embryonic mutations is displayed according to the 96 substitution classification defined by the substitution class and sequence context immediately 5' and 3' to the mutated base as described in a previous study[7]. The observed spectrum can be decomposed into just two mutational signatures (#1 and #5) that were identified previously[7].

Acknowledgements

We thank Magdalena Zernicka-Goetz at Gurdon Institute, Kevin J. Dawson at Wellcome Trust Sanger Institute and Thomas Bleazard at University of Manchester for discussion and assistance with manuscript preparation. This work was supported by the Wellcome Trust. Y.S.J is supported by EMBO long-term fellowship (LTF 1203_2012). P.J.C. is a Wellcome Trust Senior Clinical Fellow. The ICGC Breast Cancer Consortium was supported by a grant from the European Union (BASIS) and the Wellcome Trust. For the family study, Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates [CZD/16/6] and the Scottish Funding Council [HR03006].

Author Contributions

M.R.S. designed and directed the project. Y.S.J. performed overall study with bioinformatics analyses for detection of early embryonic mutations. I.M. and M.G. performed statistical testing to confirm unequal contributions of early cells and early mutation rates. L.B.A. carried out mutational signature analyses. R.R. and M.E.H. designed and directed family studies. H.R.D., M.R., S.N.-Z. performed cancer genome analyses and provided conceptual advice. M.P., A.F., C.A., N.P., S.G., and S.O.'M carried out laboratory analyses. S.M. supported clinical data analysis and curation. A.B., H.S., M.v.d.V., B.K.T.T., C.C., A.T., N.T.U., L.J.v.V., J.W.M.M., S.C., S.K., P.N.S., S.R.L., J.E.E., A-L.B-D., A.R., A.M.T., A.V. provided clinical samples and commented on the manuscript. P.J.C. supervised overall analyses. Y.S.J., I.M., M.G., L.B.A. and M.R.S. wrote the paper.

Author Information

Whole-genome sequence data have been deposited in the European Genome-Phenome Archive (EGA; https://www.ebi.ac.uk/ega/home) under overarching accession number EGAS00001001178. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.R.S (mrs@sanger.ac.uk).