SUPPLEMENTARY INFORMATION TO:

The correlation pattern of acquired copy number changes in 164 *ETV6/RUNX1*-positive childhood acute lymphoblastic leukemias

Henrik Lilljebjörn[1], Charlotte Soneson[2], Anna Andersson[1,4], Jesper Heldrup[3], Mikael Behrendtz[5], Norihiko Kawamata[6], Seishi Ogawa[7], H. Phillip Koeffler[6,8], Felix Mitelman[1], Bertil Johansson[1], Magnus Fontes[2], and Thoas Fioretos[1].

[1]Department of Clinical Genetics, University and Regional Laboratories, Skåne University Hospital [2]Centre for Mathematical Sciences, [3]Department of Pediatrics, Skåne University Hospital, Lund University, Lund, Sweden, [4]Department of Pathology, St Jude Children's Research Hospital, Memphis, TN, USA, [5]Department of Pediatrics, Linköping University Hospital, Linköping, Sweden, [6]Cedars-Sinai Medical Center, University of California, Los Angeles, CA, USA, [7]Regeneration Medicine of Hematopoiesis, University of Tokyo, School of Medicine, Tokyo, Japan and [8]National Cancer Institute of Singapore, National University of Singapore, Singapore

Supplementary Methods

High resolution genomic profiling

The 24 cases in the local data set were profiled using the Affymetrix mapping 500K array set (Affymetrix, Santa Clara, CA, USA) which consists of two 250K arrays (Sty and Nsp). Each of the two arrays provides copy number and genotype information for around 250 000 SNPs; thus, in total 500 000 SNPs, with a median distance of 2.5 kb between SNPs, are examined. In brief, 250 ng DNA was digested using either *Sty*I or *Nsp*I (New England Biolabs, Ipswich, MA, USA), respectively for the two arrays. The digested DNA was ligated to oligonucleotide adaptors using T4 DNA ligase (New England Biolabs), amplified by PCR using adaptor-specific primers, and purified using a DNA amplification cleanup kit (Clontech, Mountain View, CA, USA). Ninety micrograms of purified amplification product was fragmented using DNase I (Affymetrix) and labeled with the GeneChip DNA Labeling Reagent (Affymetrix). The labeled DNA was injected into GeneChip mapping 500K chips (Sty or Nsp) and hybridized at 49 °C for 16–18 h while the arrays where rotating at 60 rpm. The arrays were washed and stained in a Fluidics Station 450 (Affymetrix) and scanned in a GeneChip Scanner 3000 7G (Affymetrix).

Identification of copy number aberrations

For the local data set, CEL files containing raw signal intensities were exported from the GeneChip Operating Software (Affymetrix). The CEL files were imported into GeneChip Genotyping Analysis Software (Affymetrix) to produce genotype calls using the BRLMM algorithm.

The initial data analysis for the local and the two external data sets (EDS1 and EDS2) was performed in dChip (1), with all three data sets being analyzed separately. Both raw signal intensities and genotype calls were imported into the dChip software, and the raw signal intensities were normalized to a baseline level using an invariant set of probes. In dChip, the “expression value” of each

SNP was calculated from the raw signal intensities using model based expression (2) with perfect match/mismatch difference. Changes in copy number were calculated using median smoothing with a 10 SNP window. A reference copy number was calculated from all samples by trimming the 25% extreme values in both ends. To eliminate batch specific noise in the copy number data of the external data sets, copy-number calculations were performed in smaller batches based on the creation dates of the CEL-files. Circular binary segmentation was performed on the copy number data using DNAcopy (3) from the R (www.r-project.org) package Bioconductor (4). This segmentation produced three lists of copy number aberrations (CNAs), one for each data set. The lists of the two data sets of highest resolution (the local data set and EDS1) were then co-analyzed. Copy number alterations that were present in two or more samples in the two data sets were designated "recurrent changes". Copy number changes reported as polymorphisms in the database of common genetic variants (5) (http://projects.tcag.ca/variation/) were excluded since these are likely to constitute inherited copy number changes. Only regions found to harbor recurrent changes by this method were analyzed in EDS2. This data set was produced using a platform of lower resolution and for regions where EDS2 only contained between three and five probes, changes were identified by calculating the copy number using median smoothing with a 5 SNP window followed by manual inspection of the log2 copy number values and genotype information at diagnosis and remission for the probes in this region. A deletion was assumed if three or more consecutive probes had a log2 copy number below -0.3 and when the genotype information was compatible with the presence of a deletion, *i.e.,* one or more SNPs that were genotyped as heterozygous in the remission sample, were genotyped as homozygous in the leukemia sample. The sex chromosome state for external cases was inferred using the following strategy: 1) Cases with two or more copies of parts of, or the entire, X chromosome were assumed to be female (XX) if a substantial number of SNPs (> 10%) outside the psuedoautosomal regions had heterozygote calls. 2) The remaining cases were assumed to be male if the copy number of pseudoautosomal region 1 (PAR1) on Xp22.33 indicated that a Y-chromosome was present, *i.e.* the copy number of PAR1 was

higher than the copy number of the rest of the X chromosome. 3) The remaining cases, which had a difference between the copy number of PAR1 and the rest of the X chromosome that was not consistent with the presence of a Y-chromosome, were assumed to be female.

Data analysis

Branching oncogenetic trees

We used the Mtreemix software (6) to construct a branching tree model for oncogenesis (7). In this model, a causality relationship between the aberrations is assumed. First, a weighted directed graph is constructed, where the edge from aberration A to aberration B receives the weight

$$w_{AB} = \log p_{AB} - \log(p_A + p_B) - \log p_B$$

where p_{AB} is the probability of co-occurrence for the aberrations A and B, and p_A and p_B are the individual probabilities of occurrence for aberrations A and B, respectively. Note that this is not symmetric. To avoid having rare events being favored as initial vertices, one takes

$$w_{RB} = \log p_B$$

where R denotes the root event (8). Then, the maximum branching tree, *i.e.*, a tree where the sum of the weights of the included branches is as large as possible, is found using Edmond's algorithm (9).

Distance based oncogenetic trees

The basis for the distance based oncogenetic trees is a matrix containing the dissimilarities between each pair of aberrations. We calculate the dissimilarity between aberrations A and B as described by Desper *et al.* (10).

$$D(A,B) = -2\log p_{AB} + \log p_A + \log p_B$$

If A and B never co-occur in the data set, we set $p_{AB} = 0.1/N$, where N is the number of patients. We use the balanced greedy minimum evolution (GME) algorithm in the FastME software (11) to fit a

phylogenetic tree to the matrix of pairwise dissimilarities. The estimated tree has the aberrations as leaves, and the internal nodes represent unknown or "hidden" events.

Selection of nonrandom events

The method of Brodeur *et al*. (12) was used to select nonrandom events to include in the oncogenetic trees. We assume a uniform prior probability distribution for the aberrations and simulate 10,000 replicates, each with 164 patients, under this distribution. For each replicate, each aberration receives a score based on the number of occurrences in the replicate, compared to the expected number of occurrences from the prior distribution. The score is computed as

$$\frac{N_A - Np}{\sqrt{Np(1-p)}}$$

where N_A is the number of observed occurrences of aberration A in the replicate, $N = 164$ is the number of patients and p is the prior probability of occurrence for the aberration. The maximal score from each replicate is saved. The scores are then computed for each of the aberrations in the original data set, and an aberration is considered nonrandom if its score exceeds the 95th percentile of the maximal scores from the replicates. For our data set, this yields 21 nonrandom events.

Connected pair analysis

To find pairs of aberrations which co-occur either more or less often than could be expected by their individual prevalences if they were independent, we define a dissimilarity measure as follows. For two distinct aberrations A and B, let

$$D(A,B) = \frac{1}{4} - \left(\frac{N_{AB}}{N} - \frac{N_A N_B}{N^2} \right),$$

where N is the total number of samples in the data set, N_{AB} is the number of co-occurrences of A and B, N_A is the total number of occurrences of A, and N_B is the total number of occurrences of B. Then a high

value of this dissimilarity implies that *A* and *B* co-occur less frequently than could be expected by chance, and a low value indicates that *A* and *B* co-occur more frequently than expected. There is a close connection between *D* and the Pearson dissimilarity measure (*i.e.*, 1-Pearson correlation, denoted by *C*(*A*,*B*)) by

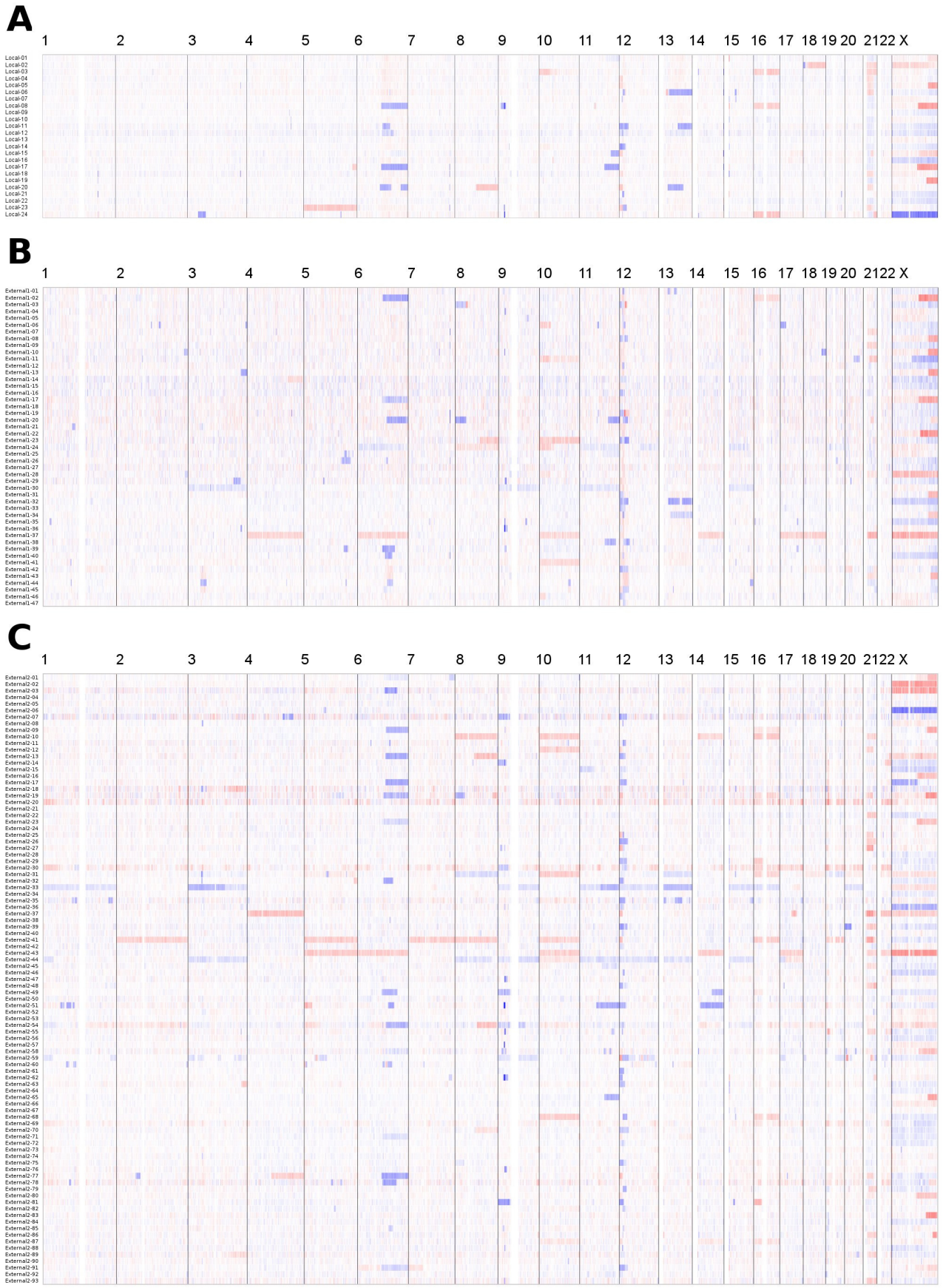
$$C(A,B) = 1 - N^2 \frac{0.25 - D(A,B)}{\sqrt{N_A(N-N_A)N_B(N-N_B)}}.$$

Hence, the correlation dissimilarity has a stronger tendency to put pairs of aberrations where the distribution is skewed at the extremes of the dissimilarity scale.

Supplementary Figure 1. Copy number changes in local and external *ETV6/RUNX1*-positive ALLs.

The log_2-transformed copy number values are plotted for each case. Blue denotes loss of material, red denotes gain, and white denotes normal copy number. Chromosome numbers are indicated above the plot, and case identifiers are indicated on the left. (A) Copy number changes in the 24 local cases (23 patients and one cell line; the cell line is designated "local-24") as determined by 500K single nucleotide polymorphism array analysis. The most common aberrations visible at this level include Xq duplications, 12p deletions and 6q deletions. (B) Copy number changes in 47 external cases published by Mullighan *et al*. (EDS1) as determined by 250K single nucleotide polymorphism array analysis. (C) Copy number changes in 93 external cases published by Kawamata *et al*. (EDS2) as determined by 50K single nucleotide polymorphism array analysis.

Supplementary Figure 1.



References

1. Lin, M., Wei, L., Sellers, W.R., Lieberfarb, M., Wong, W.H. and Li, C. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. *Bioinformatics*, **20**, 1233-1240.
2. Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 31-36.
3. Venkatraman, E.S. and Olshen, A.B. (2007) A faster circular binary segmentation algorithm for the analysis of array CGH data. *Bioinformatics*, **23**, 657-663.
4. Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.
5. Iafrate, A.J., Feuk, L., Rivera, M.N., Listewnik, M.L., Donahoe, P.K., Qi, Y., Scherer, S.W. and Lee, C. (2004) Detection of large-scale variation in the human genome. *Nat. Genet.*, **36**, 949-951.
6. Beerenwinkel, N., Rahnenführer, J., Kaiser, R., Hoffmann, D., Selbig, J. and Lengauer, T. (2005) Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics*, **21**, 2106-2107.
7. Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H. and Schäffer, A.A. (1999) Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.*, **6**, 37-51.
8. Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J. and Lengauer, T. (2005) Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.*, **12**, 584-598.
9. Edmonds, J. (1967) Optimum branching. *J Res Natl Bur Stand*, **71B**, 233-240.
10. Desper, R., Jiang, F., Kallioniemi, O.P., Moch, H., Papadimitriou, C.H. and Schäffer, A.A. (2000) Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.*, **7**, 789-803.

11. Desper, R. and Gascuel, O. (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *J. Comput. Biol.*, **9**, 687-705.
12. Brodeur, G.M., Tsiatis, A.A., Williams, D.L., Luthardt, F.W. and Green, A.A. (1982) Statistical analysis of cytogenetic abnormalities in human cancer cells. *Cancer Genet. Cytogenet.*, **7**, 137-152.