A Stem Cell Molecular Signature

Natalia B. Ivanova, John T. Dimos, Christoph Schaniel, Jason A. Hackney, Kateri A. Moore, Ihor R. Lemischka*

Mechanisms regulating self-renewal and cell fate decisions in mammalian stem cells are poorly understood. We determined global gene expression profiles for mouse and human hematopoietic stem cells and other stages of the hematopoietic hierarchy. Murine and human hematopoietic stem cells share a number of expressed gene products, which define key conserved regulatory pathways in this developmental system. Moreover, in the mouse, a portion of the genetic program of hematopoietic stem cells is shared with embryonic and neural stem cells. This overlapping set of gene products represents a molecular signature of stem cells.

Adult and embryonic stem cells (SCs) hold great promise for regenerative medicine, tissue repair, and gene therapy (*1*). Hematopoietic stem cells (HSCs) have been the most extensively studied and serve as a prototype model to define the general biological properties of mammalian SCs. Distinct developmental stages of the hematopoietic hierarchy can be identified and arranged in a hierarchical tree that begins with the long-term (LT) functional HSC. A single LT-HSC is both necessary and sufficient for life-long sustenance of the entire hematopoietic system (*2*, *3*). LT-HSCs produce less potent short-term (ST) functional HSCs, and these in turn, give rise to lineage-committed progenitor (LCP) cells. The LCP cells are directly responsible for the generation of at least 10 mature blood cell (MBC) populations. Many nonhematopoietic tissues also depend on tissue-resident SCs for their maintenance and regeneration (*4*). Totipotent embryonic stem cells (ESCs), derived from blastocysts, and neural stem cells (NSCs), derived from the germinal zones of the nervous system, are two examples of SCs that can be propagated in vitro (*5*). Because all SCs share fundamental biological properties, they may share a core set of molecular regulatory pathways. It is likely that at least some components of these regulatory pathways are preferentially expressed by SCs. We therefore attempted to define a general gene expression profile of the SC "state."

We have adopted the approach outlined in Fig. 1 that first separately identifies gene expression profiles for murine fetal and adult HSCs. These profiles are then compared to derive a shared HSC profile. This profile should include gene products that are necessary for LT hematopoietic function. We also generated gene expression profiles for human HSCs and for two murine nonhematopoietic SC populations, NSCs and ESCs. The comparison of murine with human HSCs defines evolutionarily conserved components in HSCs, whereas the comparison of hematopoietic with nonhematopoietic SCs identifies the gene products expressed in multiple SC types. The samples were processed as shown in fig. S1. Tissue or cell replicates were isolated and functionally evaluated to measure the purity of SC-containing fractions. In vitro amplified RNA probes were hybridized to Affymetrix oligonucleotide arrays. We estimate that these arrays currently allow for the monitoring of approximately 80% of HSC-related gene products (fig. S2). Arrays were scanned and processed using Affymetrix MAS 4.0 software. Genes were assigned to distinct clusters according to their expression patterns within the hematopoietic hierarchy. NSC and ESC enrichment scores were calculated to define the expression of the transcripts in these two SC populations. Bioinformatics analyses were performed for the SC-specific gene products. Details of the SC purification procedures, biological assays, and data analyses are available in supporting online material (*6*).

To translate the biological phenotypes of key hematopoietic populations into the language of gene expression, we used a series of hypothetical expression patterns that correlate with distinct, quantitatively measured biological activities present in the hematopoietic hierarchy (Fig. 2, A to C). A total of 4289 informative genes were assigned to seven clusters (Fig. 2D), characteristic of key stages of hematopoiesis, progressing from stem through progenitor to terminally differentiated cells. HSC-related clusters i to iii include many known HSC markers such as *c-Kit*, *Tie1*, *Ly-6E/Sca-1*, *Tek*, *Mpl*, *Meis1*, *Gata2*, and *Abcb1b/MDR1*. At least 72% of the above-defined HSC-related genes are also up-regulated in $CD45^{+}c\text{-}Kit^{+}Sca\text{-}1^{+}$ Hoechst 33342 side population cells (*7*). These cells have been shown to contain LT-HSCs (*8*). Furthermore, 54% of genes assigned to these clusters were previously identified through a global subtractive hybridization screen for HSC-specific gene products (*7*, *9*) (fig. S2). This demonstrates a strong correlation between HSC-specific gene sets identified by different strategies.

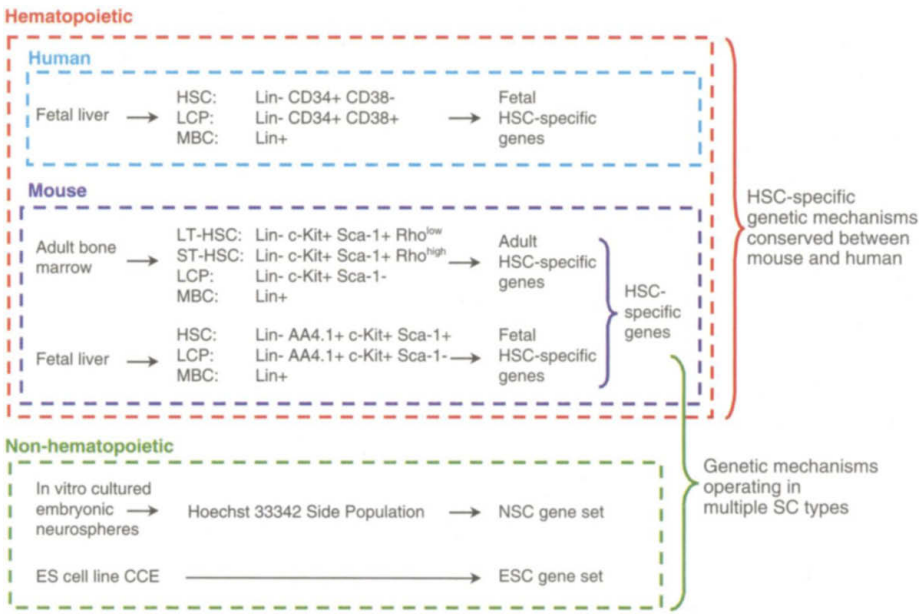


Fig. 1. Stem cell phenotypes profiled. Cells at key stages of the murine and human hematopoietic hierarchy were isolated as shown, and include LT-HSCs, ST-HSCs, LCPs, and MBCs. Nonhematopoietic SCs were cultured (ESCs) or purified (NSCs). This approach identifies three groups: genes specific for both fetal and adult murine HSCs (blue boxes), genes specific for murine and human HSCs (red box), and genes enriched in diverse SCs (green box).

Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA.

*To whom correspondence should be addressed: E-mail: ilemischka@molbio.princeton.edu

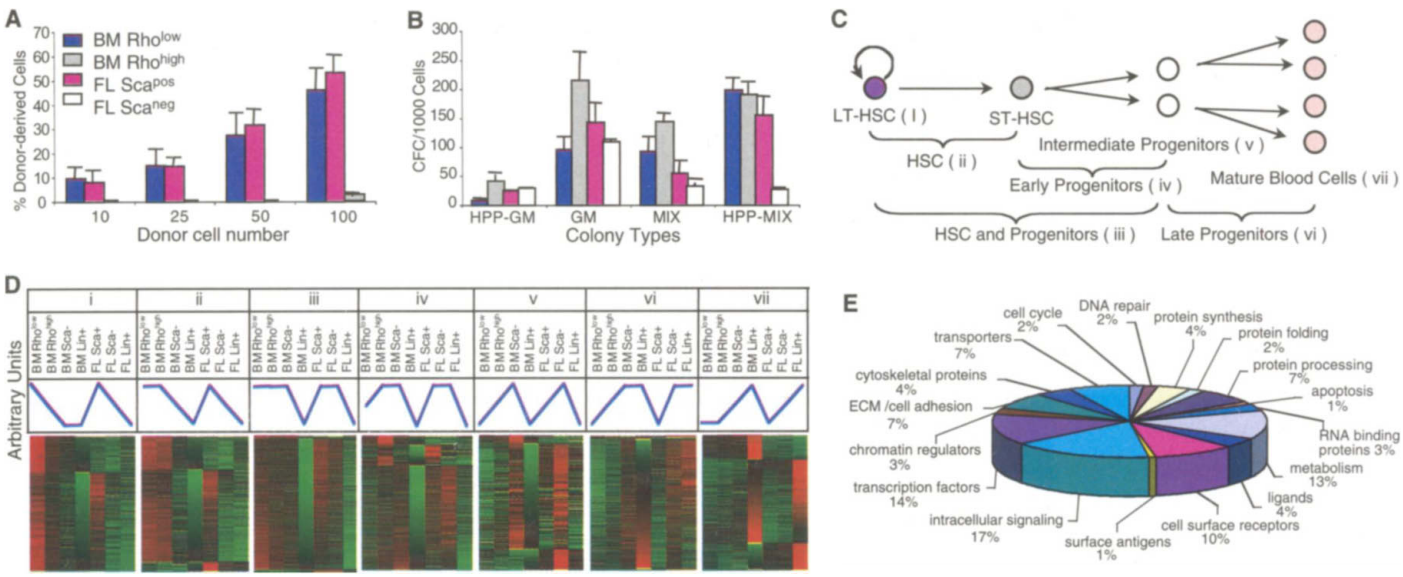
The expression specificity of 22 HSC-related genes was confirmed by quantitative reverse transcription–polymerase chain reaction (RT-PCR) (fig. S5). Gene products were grouped into categories according to their function as reported in the literature or as predicted on the basis of the presence of diagnostic protein motifs. Regulatory molecules, such as transcription factors, proteins involved in intracellular signaling, cell-surface receptors, and ligands account for 45% of the HSC-related gene-products (Fig. 2E).

We have defined genomewide transcriptional changes during early stages of hematopoietic differentiation by comparing four distinct sets of genes that are up-regulated in LT-HSCs (i), in both LT and ST-HSCs but not in LCPs (ii), in both HSCs and LCPs (iii) and, in ST-HSCs and early progenitor population (iv), respectively. The distribution of genes within these four sets across functional categories is shown in Fig. 2F. Molecules thought to be involved in cell-cell communication, such as signaling ligands, receptors, extracellular matrix, and adhesion molecules, tend to be overrepresented in the HSC-specific gene set. LT-HSC–specific ligands include *Bmp8a*, *Wnt10A*, EGF-family members *Ereg* and *Hegfl*, the angiogenesis-promoting factor *Agpt*, a ligand for the ROBO receptor family *Slit2*, and the ephrin receptor ligand *EfnB2*. These molecules may be involved in signaling between HSCs and their microenvironment. It is interesting that HSCs coexpress several ligand-receptor pairs, such as *Wnt10A/Frizzled* and *Agpt/Tek*, which suggests that HSC regulation may be partly autocrine. The complete set of HSC-related genes is presented in table S2.

ST-HSCs and early progenitors express molecules associated with the initiation of the cell cycle such as *Wee1* kinase, *Cdk4*, replication licensing factor *Mcmd*, and the critical hematopoietic proliferation protein, *Myb*. Genes involved in DNA repair and protein synthesis are also up-regulated in these compartments. This is consistent with the exit from G_0 arrest at the onset of differentiation. ST-HSCs also express a set of gene products with RNA-binding domains, which is suggestive of posttranscriptional regulation.

Hox genes are likely to play a role in HSC regulation. Four HoxA genes are expressed in different subsets of HSCs. *Hoxa5* and *Hoxa10* are specific for the LT-HSCs, *Hoxa2* is expressed in both LT and ST-HSCs, and *Hoxa9* is expressed both in HSCs and LCPs. It is noteworthy that overexpression of *Hoxa9* in murine HSCs induced stem cell expansion (*10*), whereas *Hoxa5* and *Hoxa10* perturbed their differentiation activity (*11*, *12*). In addition, *Hoxb4*, which is detected both in HSCs and LCPs, has been shown to promote specification and expansion of definitive HSCs (*13*, *14*). Fetal and adult HSCs share the key stem cell properties of self-renewal and multilineage differentiation potential. In agreement with this, comparing the gene expression profiles of fetal and adult HSCs reveals broad molecular similarities (Fig. 2G). More than 70% of all HSC-related gene-products are expressed in both fetal and adult HSCs.

We next asked whether the HSC genetic program is conserved between mouse and human. Human fetal liver $Lin^{-}CD34^{+}CD38^{-}$ cells provide long-term engraftment of non-obese diabetic immunodeficient NOD-SCID



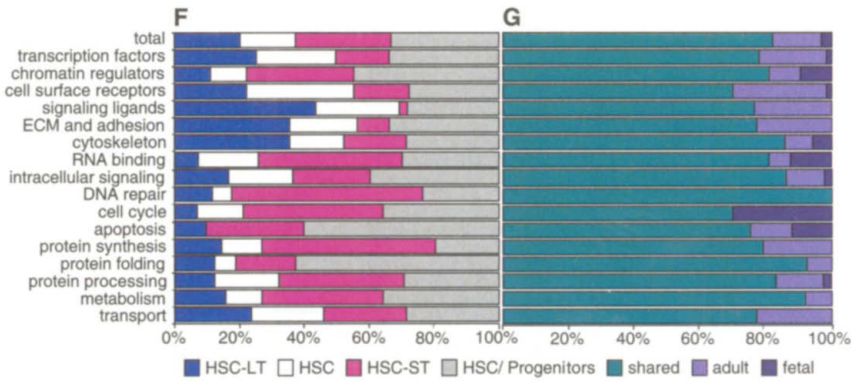
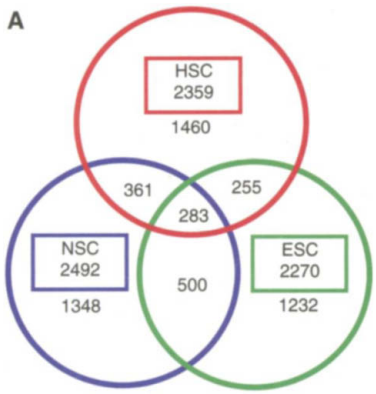


Fig. 2. Correlating biological function and gene expression. (**A**) Competitive repopulating activity of the isolated hematopoietic populations was determined (*23*). Mice were transplanted with graded doses of purified Ly5.2 fetal liver (FL) or bone marrow (BM) SCs, mixed with 2×10^5 Ly5.1 whole BM cells. Ly5.2 peripheral blood content at 6 months is shown. The repopulating stem cell frequency in these purified populations is 1 in 10 to 20 cells for both FL and BM SCs. (**B**) The number of colony-forming cells (CFCs) in the isolated stem and progenitor cell populations was determined. Colonies were scored as high proliferative potential–granulocyte macrophage (HPP-GM), GM, MIX (three or more lineages: GM, megakaryocyte, erythrocyte), and HPP-MIX. (**C**) The hematopoietic hierarchy subgrouped into different stem and progenitor populations and (**D**) their corresponding expression clusters (i to vii). Individual genes were assigned to expression clusters as described (*6*). Relative expression levels are displayed by red (highest) to green (lowest) coloration. Predicted cellular roles of identified HSC-specific gene products: (**E**) distribution within the HSC profile for gene products with known or putative functions, (**F**) distribution of the annotated gene-products between HSC subtypes, and (**G**) between fetal and adult HSCs.

mice and, therefore, are functionally similar to murine LT-HSCs (*15*). Human gene products with an increase in expression of at least twofold in HSCs compared with MBCs were defined as HSC-enriched. Mouse-human homologous pairs were identified by direct sequence comparison of expressed sequence tag (EST) assemblies as described (*6*). We found 822 human homologs for murine HSC-related genes that are expressed in fetal liver (Database S3). Of these, 322 (39%) were enriched in human fetal HSCs. The probability of observing such an overlap by chance as estimated using hypergeometrical distribution (*6*, *16*) is extremely low ($P = 10^{-11}$). These genes likely represent the conserved molecular components expressed in HSCs. Homologous gene products expressed in the LT-HSC subset are listed in Table 1. The remaining homologous pairs did not show coordinate expression. This may reflect technical difficulties in purifying homogeneous HSC fractions. Alternatively, related but not identical populations may function as HSCs in different organisms.

To establish the gene expression profile common for diverse types of SCs, we performed analyses of ESCs and NSCs. Gene products with an increase in expression of at least twofold compared with both fetal and adult MBCs were defined as ESC/NSC-enriched. Correct detection of ESC- and NSC-enriched genes was verified by comparison with published data sets (*17*, *18*) and are presented in tables S3 and S4.

ESC/NSC-enriched gene sets were compared with each hematopoietic cluster. These



B

Hematopoietic Clusters	NSC ESC	NSC Only	ESC Only	Hemato Only	All Genes
LT-HSC (i)	52	94	38	406	590
HSC (ii)	43	67	32	342	484
HSC & Progenitors (iii)	146	150	114	476	886
ST-HSC & Early Progenitors (iv)	42	50	71	236	399
Total HSC-related	283	361	255	1460	2359
Intermediate Progenitors (v)	13	16	40	149	218
Late Progenitors (vi)	5	12	25	663	705
Mature Blood Cells (vii)	2	8	8	623	641

Fig. 3. Overlapping gene expression in diverse murine SCs. **(A)** Venn diagram detailing shared and distinct gene expression among NSCs, ESCs, and HSCs. **(B)** A summary of the number of different genes expressed in diverse stem cell compartments in relation to each other and compared with the above defined hematopoietic clusters. A complete list of HSC-related genes also enriched in NSCs and/or ESCs is presented in Database S4.

Table 1. Select known mouse-human homologs expressed in LT-HSC subset. The complete list of homologous pairs is presented in Database S3. FC, fetal cells, GPCR, G protein–coupled receptor; LDL, low density lipoprotein; MHC, major histocompatibility complex; TF, transcription factor; UTR, untranslated region.

Gene Name	Mouse GenBank ID	Human GenBank ID	Mouse FC	Human FC	Annotation
Ches1	AW046392	U68723	3.3	2.8	Checkpoint suppressor 1, DNA damage, cell-cycle arrest
SREC	AA986099	D86864	10	3.6	Acetyl LDL scavenger receptor
Blr1	AI608284	X68149	2.8	3.6	Burkitt lymphoma–associated chemokine GPCR
Procr	L39017	L35545	39.3	3.1	Endothelial cell protein C receptor
Fzd4	U43317	AI927489	9.2	2.4	Frizzled-like GPCR (Wnt receptor)
Igf1r	AF056187	X04434	4.6	2.1	Insulin-like growth factor I receptor
Mtap7	Y15197	X73882	3.3	6.3	Microtubule-associated protein
MYO5C	AW214321	AA195002	9	6.4	Myosin 5 motif
Pclo	Y19186	AB011131	7.8	2	Presynaptic cytomatrix protein
Sparcl1	AV347505	X86693	4.6	3.1	SPARC-like protein
Ocln	AW209088	U49184	6.3	2	Tight junction component
Jcam2	AI853724	AI199779	15.2	3	Tight junction component
Jcam3	AI850297	AA149644	11.1	8.7	Tight junction component
Mpdz	AV244715	AF093419	18.4	5.2	Multiple PDZ domains, interacts with GPCRs
Nbea	AI154580	AI052524	29.3	3.2	Protein kinase A regulator
SCOP	AI836256	AB011178	3.7	2.7	Protein phosphatase 2C domain
Ptpn21	D37801	X79510	32.4	3.8	Protein tyrosine phosphatase, nonreceptor type
Ndr2	AV349686	AI201607	14.7	2.1	Regulated by N-*myc*
Rras	M21019	AI201108	9.8	2.7	R-*ras* oncogene
Agpt	U83509	U83508	9.8	20.9	Angiopoietin-1, binds TIE-2/Tek receptor
Efnb2	U30244	AI765533	39.4	2.9	Ephrin B2, Eph receptor ligand
Rbp1	X60367	M11433	49.8	3.8	Involved in metabolism of retinoids
Aldh2	AV329607	X05409	21.9	4.1	Mitochondrial aldehyde dehydrogenase
Fkbp7	AF040252	AI271550	3.3	2.9	Peptidyl-prolyl *cis-trans* isomerase
Smpd1	AV347445	M81780	13.2	2.1	Lysosomal sphingomyelin phosphodiesterase
Tapbp	AV361189	AA767887	7.9	2.3	MHC-like antigen-processing transporter
Nnp1	AV260279	AI860822	2.5	5	Nucleolar protein 52–like
Htf9c	AV325777	AW007779	6.5	2	RNA recognition motif–containing protein
Elavl4	AV241912	AA102788	13.9	6.4	Uridylate-rich UTR binding
Tcf3	AJ223069	AI916838	5.6	2	General immunoglobulin TF-3
Pphn	AW123178	M95585	33.8	201.4	Hepatic leukemia factor implicated in apoptosis inhibition
Hoxa5	Y00208	AC004080	3	4.1	Up-regulates p53 and progesterone receptor expression
P2rx4	AF089751	U83993	9	2.9	Purinergic receptor ligand–gated ion channel
Slc12a2	U13174	N56950	10.6	2.9	Sodium-potassium-chloride cotransporter

results are summarized in Fig. 3. Gene products enriched in all three SC types belong to a variety of functional categories. Several identified gene products have been previously implicated in the regulation of different types of SCs. Transcription factors *Edr1* and *Tcf3* have been shown to sustain the activity of HSCs (*19*) and epidermal SCs (*20*), respectively, whereas *EfnB2* and *Hes1* have been implicated in control of NSC proliferation (*21*, *22*). Analyses of EST collections indicate that many of the HSC-ESC- and NSC-enriched genes are also expressed in other tissues (*7*). This may suggest more general functional roles in a broader array of SC populations.

In summary, we have determined the molecular similarities and differences among five distinct SC populations, specifically, human fetal HSCs, murine fetal and adult HSCs, NSCs, and ESCs. The similarities define a common SC genetic program or SC molecular signature. It is likely that hallmark properties shared by all SCs, such as the ability to balance self-renewal and differentiation, will be governed by shared molecular mechanisms. As such, numerous components of these molecular mechanisms are likely to be contained within the SC molecular signature presented here.

References and Notes

1. I. L. Weissman, *Science* **287**, 1442 (2000).
2. C. T. Jordan, I. R. Lemischka, *Genes Dev.* **4**, 220 (1990).
3. M. Osawa, K. Hanada, H. Hamada, H. Nakauchi, *Science* **273**, 242 (1996).
4. E. Fuchs, J. A. Segre, *Cell* **100**, 143 (2000).
5. I. L. Weissman, D. J. Anderson, F. Gage, *Annu. Rev. Cell Dev. Biol.* **17**, 387 (2001).
6. Material and Methods are available as supporting online material on *Science* Online.
7. N. B. Ivanova, K. A. Moore, I. R. Lemischka, unpublished observations.
8. M. A. Goodell, K. Brose, G. Paradis, A. S. Conner, R. C. Mulligan, *J. Exp. Med.* **183**, 1797 (1996).
9. R. L. Phillips *et al.*, *Science* **288**, 1635 (2000).
10. U. Thorsteinsdottir *et al.*, *Blood* **99**, 121 (2002).
11. G. M. Crooks *et al.*, *Blood* **94**, 519 (1999).
12. C. Buske *et al.*, *Blood* **97**, 2286 (2001).
13. J. Antonchuk, G. Sauvageau, R. K. Humphries, *Cell* **109**, 39 (2002).
14. M. Kyba, R. C. Perlingeiro, G. Q. Daley, *Cell* **109**, 29 (2002).
15. G. Guenechea, O. I. Gan, C. Dorrell, J. E. Dick, *Nature Immunol.* **2**, 75 (2001).
16. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, G. M. Church, *Nature Genet.* **22**, 281 (1999).
17. D. L. Kelly, A. Rizzino, *Mol. Reprod. Dev.* **56**, 113 (2000).
18. D. H. Geschwind *et al.*, *Neuron* **29**, 325 (2001).
19. H. Ohta *et al.*, *J. Exp. Med.* **195**, 759 (2002).
20. B. J. Merrill, U. Gat, R. DasGupta, E. Fuchs, *Genes Dev.* **15**, 1688 (2001).
21. J. C. Conover *et al.*, *Nature Neurosci.* **3**, 1091 (2000).
22. T. Ohtsuka, M. Sakamoto, F. Guillemot, R. Kageyama, *J. Biol. Chem.* **276**, 30467 (2001).
23. D. E. Harrison, C. T. Jordan, R. K. Zhong, C. M. Astle, *Exp. Hematol.* **21**, 206 (1993).

24. We thank C. Jordan for providing the human hematopoietic samples, A. Beavis for expert flow cytometry, and T. Doniger and M. Pritsker for assistance with bioinformatics. We also thank N. Stahl and F. Santori for critically reviewing the manuscript. This work was supported by grants from the NIH DK54493 and DK42989 (to I.R.L.). Additional support was provided by ImClone Systems, Inc., New York.

Supporting Online Material

www.sciencemag.org/cgi/content/full/1073823/DC1
Materials and Methods
Fig. S1 to S5
Tables S1 to S4
Databases (Excel files) 1 to 4

10 May 2002; accepted 3 September 2002
Published online 12 September 2002;
10.1126/science.1073823
Include this information when citing this paper.

Signal-Driven Computations in Speech Processing

Marcela Peña,[1] Luca L. Bonatti,[1,2] Marina Nespor,[3] Jacques Mehler[1,4]*

Learning a language requires both statistical computations to identify words in speech and algebraic-like computations to discover higher level (grammatical) structure. Here we show that these computations can be influenced by subtle cues in the speech signal. After a short familiarization to a continuous speech stream, adult listeners are able to segment it using powerful statistics, but they fail to extract the structural regularities included in the stream even when the familiarization is greatly extended. With the introduction of subliminal segmentation cues, however, these regularities can be rapidly captured.

To learn an unknown language, listeners must segment connected speech into constituents and discover how words are organized. When adults try to cope with an unknown language or when infants learn their native language, they do so by listening to speech before they know either the words or the grammatical system of that language, and without receiving explicit instruction. To extract words as well as their organization from the speech stream, infants and adults must possess efficient computational procedures.

Several solutions have been proposed to account for speech segmentation (*1*, *2*). In particular, some investigators (*3*–*5*) have shown that adults and 8-month-old infants confronted with unfamiliar concatenated artificial speech tend to infer word boundaries at loci where the transitional probability between two adjacent syllables drops. That is, word boundaries are inferred between two syllables that rarely appear in sequence and not between two syllables that always appear together (*6*). Saffran *et al.* (*5*) demonstrated that participants exposed for several minutes to continuous speech judge trisyllables delimited by dips in transitional probability as being more familiar than trisyllables enclosing a transitional probability dip. Other studies have helped establish the importance of statistics in parsing speech as well as nonspeech sequences: adults can take advantage of statistics to segment speech streams, sequences of tones (*7*), and sequences of visual stimuli (*8*–*10*), among other types of sequences.

As to the mechanisms responsible for the extraction of structural information, little is known. In one study (*11*), 7-month-old infants behaved as if they had inferred a rule after having been familiarized with a large number of trisyllabic items consistent with it. After familiarization, infants were presented with previously unheard items, and they behaved differently according to whether or not the items conformed to the rule. This result was observed using segmented strings of items composed of three separate consonant-vowel syllables (*12*). This suggests that infants tend to extract rule-like regularities, at least when they process a corpus of clearly delimited items. This study emphasizes the specific computational abilities that favor the discovery of the structural properties of a corpus. Conceivably, in the absence of such abilities, language would be impossible to acquire.

Assessing the scope and limits of statistical and structural computations for learning words and grammar in language remains an elusive problem. One reason is that the methodologies and stimuli used in the above-cited studies are sufficiently different that the relative importance of the two underlying mechanisms cannot be directly compared. The aim of our study is to explore, by means of easily comparable experimental situations, what such mechanisms accomplish and when precisely they operate in language processing. To this purpose, building on a suggestion by Newport and Aslin (*13*), we explore whether participants can segment a stream of speech by means of nonadjacent transition probabil-

[1]International School for Advanced Studies, Trieste, Italy. [2]University of Paris VIII at Saint Denis, France. [3]University of Ferrara, Italy. [4]Laboratoire de Science Cognitive et Psycholinguistique, CNRS and EHESS, Paris, France.

*To whom correspondence should be addressed. E-mail: mehler@sissa.it