Inimese mittekodeeriva DNA suuremahuline klasterdamine

JC 6.10.2004

Inimese DNA

- Vähemalt 5% genoomist on konserveerunud
- Kodeerivad eksonid 1.5% (2% koos UTR)
- Ülejäänud 3-3.5%
  - Transkriptsioonisaidid
  - Splaissingu saidid
  - RNA geenid
  - micro-RNA
  - Replikatsiooni originid
  - ...

Senini

- Andmebaasid mittekodeerivate funktsionaalsete regioonide jaoks
- Inimese-hiire võrdlev genoomika
- Homoloogsete valkude leidmine

Eesmärgid

- Leida kõrgelt konserveerunud regioonid inimese genoomis võrdluses hiire ja roti genoomidega
- Ennustada võimalike funktsioone leitud regioonidele

Meetodid

- Inimese genoomi regioonide võrdlus sünteeniliste homoloogidega hiire genoomis
- Klasterdamine graafide abil
- G= (V,E)
  - Tipud V = inimese konserveerunud regioonid
  - Servad E = regioonide paar millel oluline järjestuse sarnasus inimese genoomi siseselt

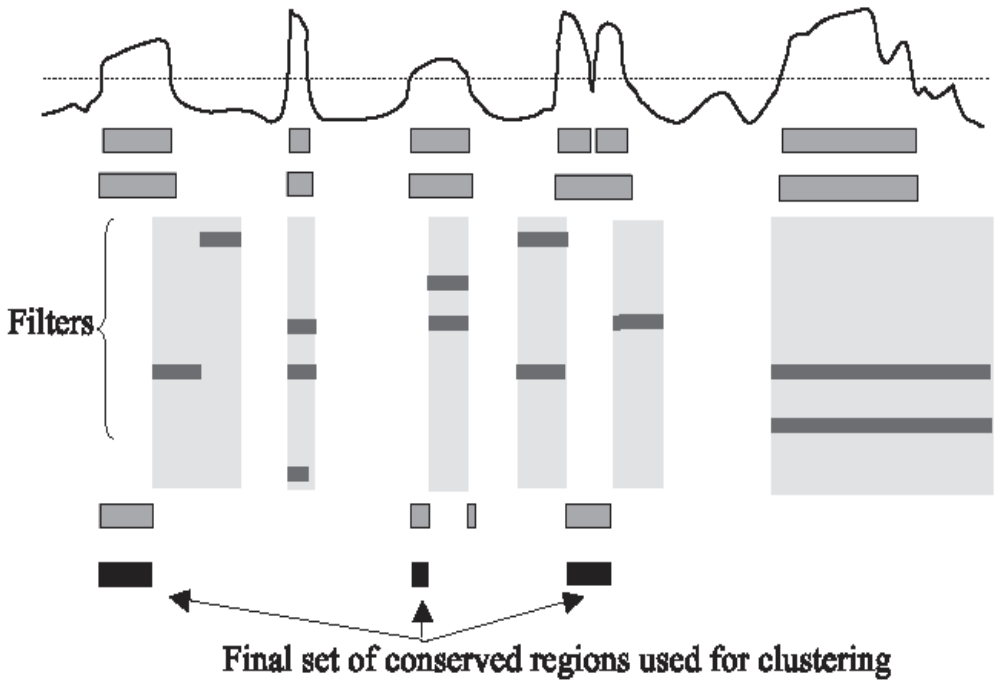
Konserveerunud elemendid

- Regioonid mis on kõrgelt konserveerunud võrreldes hiire ja roti ortoloogidega joondati 3-se mitmese joondamisega
- Joondust skanneeriti 50bp libiseva aknaga, igas aknas arvutati p-value konserveerumise astmele

Valitud regioonid

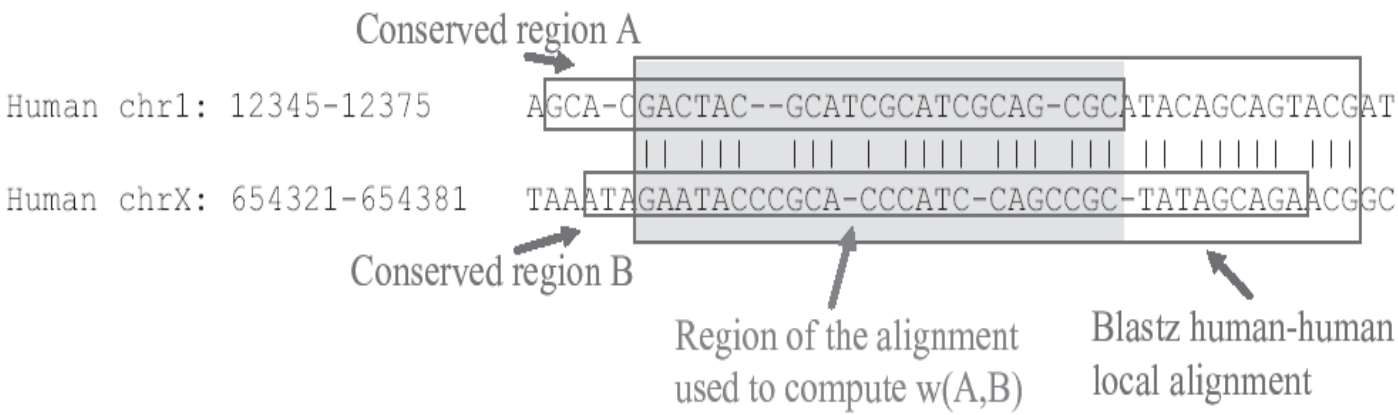
- 5% inimese genoomist hõlmatud
- 140bp keskmine pikkus
- 74% kodeerivatest eksonitest valikus moodustades 13% (17%) valitud regioonidest
- Pikendati iga regiooni 10bp võrra nii up-kui downstream suunal
- Rakendati filtreid

Filtrid



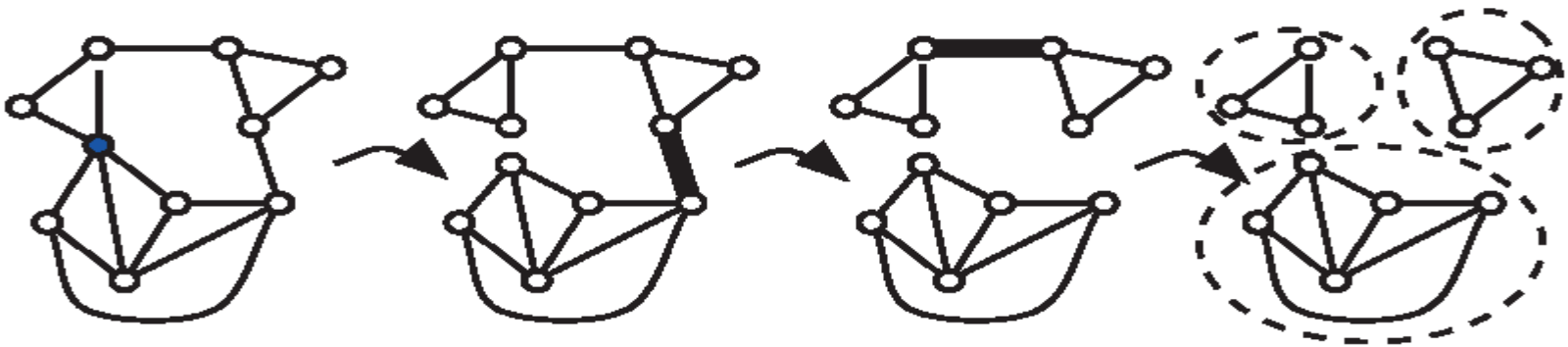
Human-mouse-rat sequence conservation	Fraction of genome	Number of regions
Top ~5%	5.11%	1055823
Extend/merge	5.835%	969857
Remove repeats	5.393%	959820
Remove coding exons	4.541%	1074181
Remove cds-like	4.239%	1084945
Remove non-syntenic	4.086%	1072148
Remove seg. dups.	4.006%	1043450
Remove pseudogenes	4.001%	1042608
Remove <50bp	3.727%	699647

- Sarnasusgraaf
- Blastz
- (u,v) ühendatakse servaga kui on järjestikune ≥15bp joondus ja servale antakse kaal w(u,v) = s(u,v)



- 699647 tippu
- 6% ühendatud
- 29349 regiooni
- 8333 ühendatud komponenti
  - 1446 vähemal 3-tipulised
  - 257 vähemalt 10-tipulised
- Suurim 823 tippu, 1673 serva

Servade kõrvaldamine



Tulemused

- 12027 tihedat klastrit
  - 2-105 regiooni
  - 296 -> 5 või enam regiooni
  - 84 -> 10 või enam regiooni
  - tipu aste 4-10
  - 6.1 keskmine tipu aste
  - 18734 inimese genoomi regiooni
  - 225 aluspaari

Leitud klastrite võrdlus

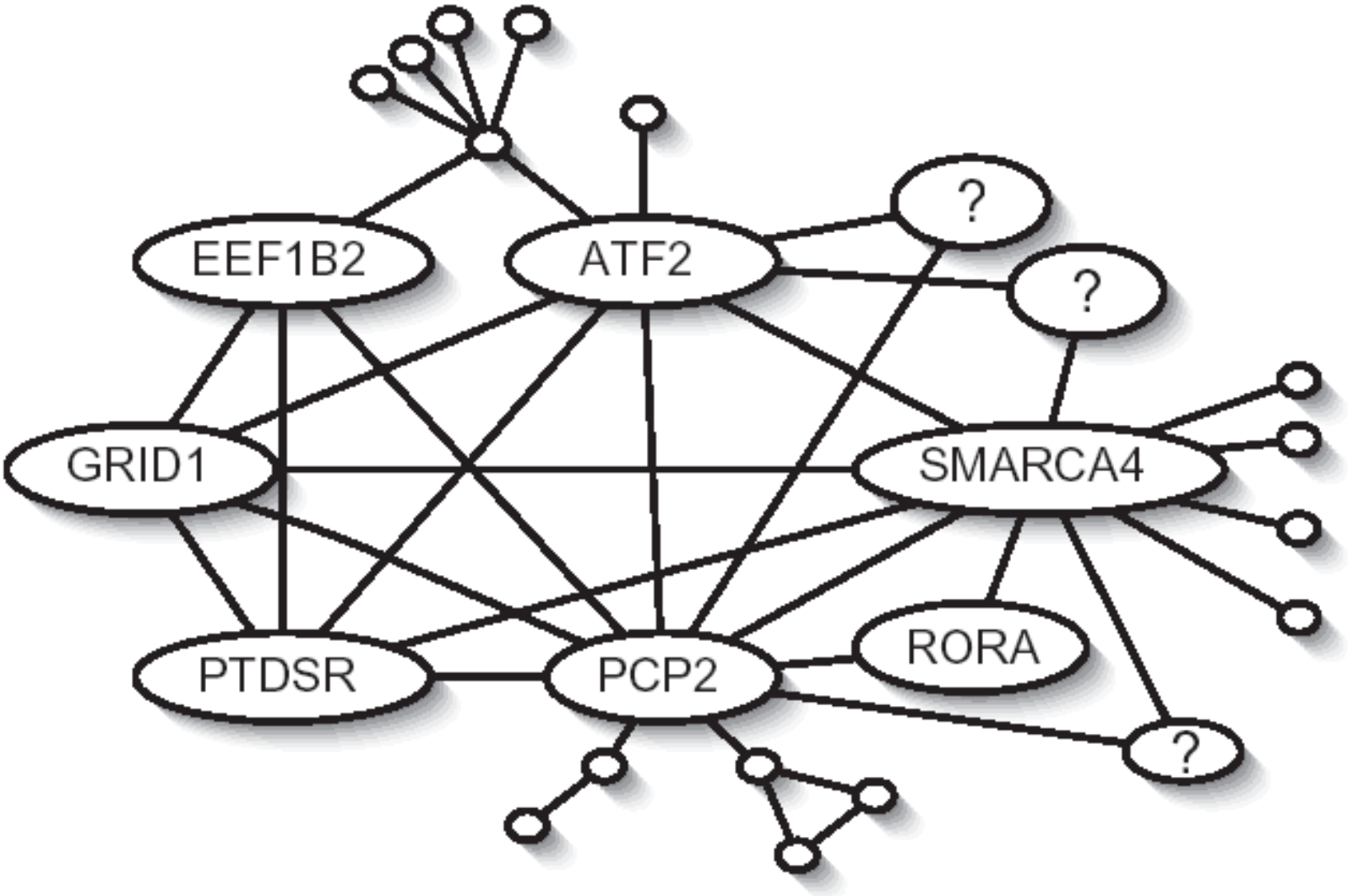
- Eeldatakse et klastri liikmetel jagavad ühist funktsionaalsust
  - Ühe liikme annotatsiooni saab võrrelda teiste liikmete omadega
  - Statistiline üleesindatus mingi omaduse suhtes võib viidata võimalikule ühisele funktsioonile

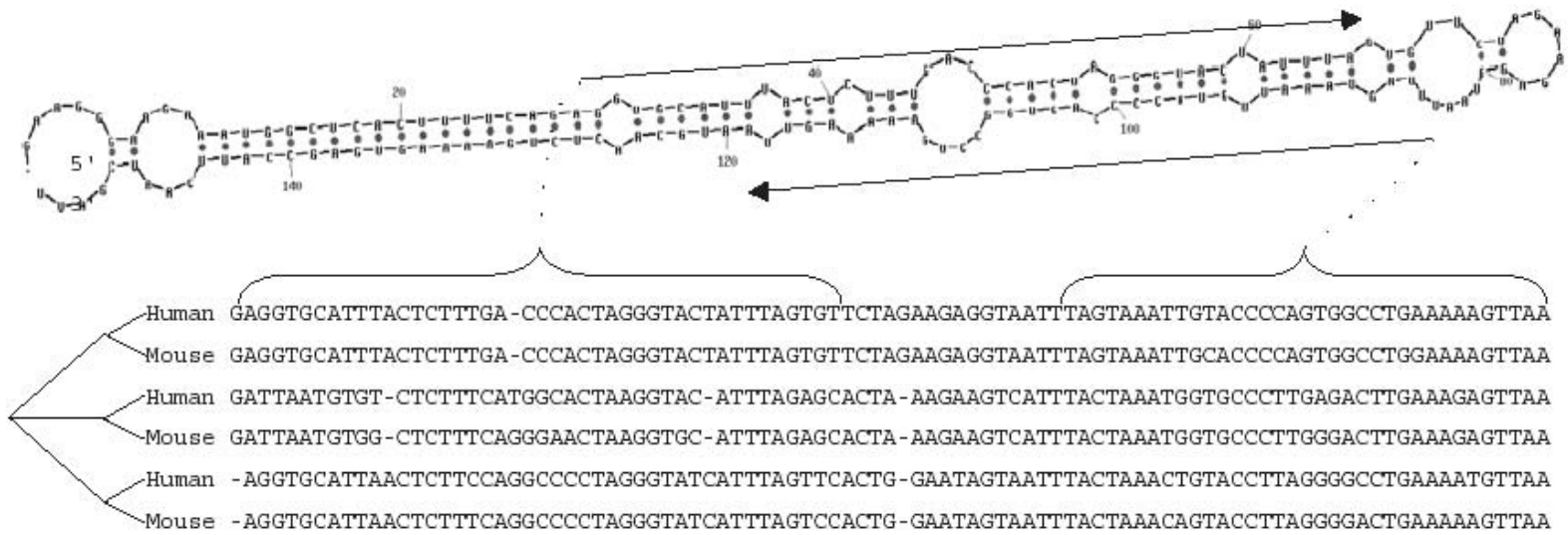
Funktsiooni otsingud

- Genoomne lokalisatsioon
- Assotsiatsioon tuntud geenidega
- Kodeerimispotentsiaal
- Transkriptsiooni tõendid
- Mittekodeeriv RNA
- Ennustatav RNA sekundaarstruktuur
- Konserveerumine kaugetel liikidel

Attribute	Cluster ID	#v	#Att	p-value	Comment
RNA genes	5390.1	6	6	9.7e − 22	Hu-U71b snoRNAs
	2483.22	9	4	1.2e − 12	miRNA mir-154. Also detected by RNA sec. struct. p-value screening
	41 others			<1.6e − 08	various RNAs and miRNAs
Chicken conservation	14.381	59	38	3.7e − 13	No conservation in fugu
	156.175	16	15	6.3e − 10	Many matches to chicken EST
	1730.12	13	11	1.4e − 6	Five regions have coding potential (p-value 4.9e−4)
	2003.3	19	15	8.1e − 8	Ten regions have coding potential (p-value 1.8e−8) and 8 regions have RNA secondary structure (p-value 7.2e−13)
Fugu conservation	4415.3	5	5	7.9e − 11	Just 5′ of exons of SCNxA gene family (p-value 8.4e−6), all are conserved in chicken (p-value 3.7e−4)
	4290.2	4	4	8.3e − 9	3′ end of 5′-UTR of histone H1 family
	4787.3	4	4	8.3e − 9	Downstream of alt. splices exons of the NEB gene
	5602.2	4	4	8.3e − 9	All are predicted genes with EST evidence
	855.1	4	4	8.3e − 9	All have strong RNA sec. str (p-value 1e−8)
	24 others			<8.6e − 07	
ESTs	652.29	10,21	6	9.7e − 7	Six sites are <1 kb downstream of exons of various genes (Fig. 3B).
Upstream	6137.8	11	10	2.6e − 17	5′ of genes of the ALEX family. Many other clusters are associated with the same family
	6895.5	5	4	4.4e − 7	Just 5′ of genes of the PCDHB family
	1848.5	4	4	4.4e − 7	Just 5′ of genes of the KRTHA family
	4982.2	5	5	2.8e − 7	5′-UTR of genes of the SCNxA family. Many other clusters are associated with the same family
	5105.1	5	4	4.4e − 7	5′-UTR of genes of the GRYD family
	4 other clusters			<5.2e − 6	Various gene families
1 kb intron flanks	6898.2	12	11	7.5e − 11	Downstream of alternative first exons of PCDHG family Many other clusters are associated with the same family
	4969.6	12	9	1.2e − 7	Upstream of repetitive exons of TTN
Gene predictions	7708.1	15	15	1.8e − 19	Consecutive regions contained in a 12 kb ORF upstream of c2orf16
	5011.6	5	5	5.6e − 7	Consecutive regions contained in a 5 kb ORF upstream of AK126051
	3089.3	5	5	3.1e − 8	Similar to collagen alpha 3 VI chain precursor
RNA sec. struct.	652.45	25	13	4.6e − 20	8 regions overlap gene predictions
	221.127	12	9	2.1e − 16	See Fig. 4
	50 others			<1e − 6	
Go/InterPro annotation	631	18	15	1e − 18/1e − 28	Mostly intronic, to various homeobox transcription factors

#v is the number of vertices in the cluster. #Att describes the number of cluster members that have a given attribute (also see Supplementary Material).


EEF1B2
ATF2
?
?
GRID1
SMARCA4
PTDSR
PCP2
RORA
?


5'
3'
20
40
60
80
100
120
140
Human GAGGTGCATTTACTCTTTGA-CCCACTAGGGTACTATTTAGTGTTCTAGAAGAGGTAATTTAGTAAATTGTACCCCAGTGGCCTGAAAAAGTTAA
Mouse GAGGTGCATTTACTCTTTGA-CCCACTAGGGTACTATTTAGTGTTCTAGAAGAGGTAATTTAGTAAATTGCACCCCAGTGGCCTGGAAAAGTTAA
Human GATTAATGTGT-CTCTTTCATGGCACTAAGGTAC-ATTTAGAGCACTA-AAGAAGTCATTTACTAAATGGTGCCCTTGAGACTTGAAAGAGTTAA
Mouse GATTAATGTGG-CTCTTTCAGGGAACTAAGGTGC-ATTTAGAGCACTA-AAGAAGTCATTTACTAAATGGTGCCCTTGGGACTTGAAAGAGTTAA
Human -AGGTGCATTAACTCTTCCAGGCCCCTAGGGTATCATTTAGTTCACTG-GAATAGTAATTTACTAAACTGTACCTTAGGGGCCTGAAAATGTTAA
Mouse -AGGTGCATTAACTCTTTCAGGCCCCTAGGGTATCATTTAGTCCACTG-GAATAGTAATTTACTAAACAGTACCTTAGGGGACTGAAAAAGTTAA

Edasiseks lugemiseks

- Identification and Characterization of Multi-Species Conserved Sequences (Margulies, Blanchette, Haussler, Green)